UNIVERSITY OF CALIFORNIA

Los Angeles

Essays in Macroeconomic Dynamics

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Economics

by

Flavien Erwin Alain Moreau

2019

ABSTRACT OF THE DISSERTATION

Essays in Macroeconomic Dynamics

by

Flavien Erwin Alain Moreau
Doctor of Philosophy in Economics
University of California, Los Angeles, 2019
Professor Andrew Granger Atkeson, Chair

This dissertation focuses on the macroeconomic implications of firms' behavior. How do capital and labor substitute for each other? How large are the inefficiencies created by anti-competitive behavior? How does the distribution of the size of firms interact with merger regulations?

In the first chapter, I investigate the question of the substitutability of labor and capital at the firm-level and then in the whole economy. Using a novel empirical strategy and comprehensive administrative data. I find that the amount of substitution between capital and labor is actually fairly limited.

In the second chapter, I study the distortions created by anti-competitive behavior in an oligopolistic setting. I find that the direct negative welfare impact of cartels is amplified by an umbrella pricing effect, whereby firms outside of the cartel also raise their prices.

In the third chapter, I trace the implications of antitrust policies on the firm size distribution. I find that under threshold-based rules, all mergers above a certain size might need to be blocked in order for the size distribution to stabilize.

Finally, the techniques used to study the dynamics of firms can fruitfully be applied to other areas. In particular, in the last chapter, jointly written with Adriana Lleras-Muney, we use a dynamic model to explore the evolution of mortality rates.

The dissertation of Flavien Erwin Alain Moreau is approved.

Pierre-Olivier Weil

Andrea Lynn Eisfeldt

Shuyang Sheng

Andrew Granger Atkeson, Committee Chair

University of California, Los Angeles

2019

To my parents

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I thank my advisor Andy Atkeson for his guidance and numerous insightful discussions.
I also thank Pierre-Olivier Weil, Shuyang Sheng, and Andrea Eisfeldt, as well as Francois Geerolf, Adriana Lleras-Muney, and John Asker for detailed comments and support.
Lee Ohanian, David Baqaee, Jon Vogel, Ariel Burstein, Pablo Fajgelbaum, Saki Bigio, Ryan Martin, Vladimir Pecheu, Yanos Zylberberg, and participants of various seminars provided helpful comments.
I gratefully acknowledge the hospitality of Sciences-Po Paris and of Aix-Marseille School of Economics. In particular, I thank Thomas Chaney for helping me getting access to the French confidential administrative data. Juliana Smith, Chiara Paz, Sandrine Le Goff, and Sandrine Guillerm provided competent, timely, and generous assistance.
I also want to thank employees at Ministry of Finance and at the French National Statistical Agency, who have helped me better understand the data. Finally, for providing access to admnistrative data, I am grateful to the Centre d'accès sécurisé aux données (CASD), which is supported by a public grant overseen by the French National Research Agency (ANR), as part of the «Investissements d'avenir» program (reference : ANR-10-EQPX-17).// I was privileged to be hosted by the Camargo Foundation in Spring 2019, when this manuscript was composed.
Finally, I am deeply indebted to Madeline Woker for her continuous and unending support and without which these thesis would not have been written.

VITA

2008–2012 M.Sc. Engineering, Ecoles des Ponts ParisTech.

2011–2012 M.Sc. Econometrics and Mathematical Economics, London School of Economics and Political Sciences.

2013-2015 M.A. Economics, UCLA.

CHAPTER 1

Inferring Capital-Labor Substitution from Firm-level Distortions

I propose a novel identification strategy to estimate the elasticity of substitution between capital and labor at the firm level, taking advantage of a set of size-dependent policies. These policies create a break in the size distribution of French firms by employment. To the extent that firms around the threshold distort their capital-labor ratio, the elasticity of substitution can be recovered from firm-level distorted factor ratios and the amount of bunching. On the other hand, because firms have an incentive to under-report their size, part of the apparent distortions is only due to evasion. I derive closed-form formulas that link the elasticity to observed distortions at the firm level. I then aggregate these micro elasticities to analyze capital-labor substitution in the aggregate economy. I find that firm-level distortions are small, and therefore that the micro- elasticities are small and close to 0.1. I obtain an aggregate elasticity of 0.3, about half as small as the values usually assumed. Moreover, the discrepancies between self-reported employment and a measure of employment made by the administration provide a direct evidence that evasion accounts for a large fraction of the observed distortions.

"Machinery and labour are in constant competition, and the former can frequently not be employed until labour rises. (...) The demand for labour will continue to increase with an increase of capital, but not in proportion to its increase; the ratio will necessarily be a diminishing ratio."

Ricardo, 1817, *On The Principles of Political Economy and Taxation*, ch. 31 On Machinery.

1.1 Introduction

Quantitative assessments of the phenomenon described by Ricardo 1817 rely on a measure of the elasticity of substitution between capital and labor. This parameter plays a central role in macro-economic analysis. This elasticity not only governs the distribution of the labor share, but also determines the impact of investment and innovation on long-run growth, or the incidence of corporate taxation. While the bulk of the evidence suggests that this aggregate elasticity ranges around 0.60.7, a firm consensus remains elusive.[1] Whether this elasticity is greater or smaller than one, however, entails completely opposite economic implications.[2] For instance, in a widely cited article, Karabarbounis and Neiman 2014 estimate an aggregate elasticity of 1.25, with which they account for the reduction of the labor share with a decrease of capital prices. With an elasticity smaller than 1, a decrease of capital prices would have *increased* the labor share.

This paper proposes the first quasi-natural experimental design to estimate this key elasticity, using the distortions that several size-dependent regulatory thresholds create on the size distribution of French firms' and on their choices of inputs.[3] These distortions arise from firms' reactions to a combination of administrative and organizational regulations that

[1]See Antras 2004, Chirinko 2008, or León-Ledesma, McAdam, and Willman 2010 for recent surveys of elasticity estimates.

[2]This important point was recognized by Hicks 1932 in the first analytical study of the elasticity of substitution between capital and labor.

[3]In contemporaneous and independent research, Benzarti and Harju 2018 follow an approach similar to mine, but using a regulatory threshold in Finland which is based on the amount of capital depreciation. Interestingly, they find micro elasticities that are exactly 0.

become mandatory once a firm reaches a certain number of employees. I consider three main thresholds, standing respectively at 10, 20, and 50 employees. Standard profit maximization predicts that in the face of discontinuities in their cost functions, some firms will prefer to bunch before the threshold instead of growing past it. Unless their production functions are Leontief, that is, with fixed factors, it is optimal for these firms to substitute capital for labor in order to avoid incurring the regulation costs. As a consequence, the extent to which this substitution actually occurs empirically informs us about the elasticity of substitution at the firm-level.

The key intuition of this paper is to jointly exploit two moments of the data – the amount of bunching in firms' labor size and the distortion in capital intensity – in order to recover the elasticity of substitution and the cost of the policy. A strong response to the regulation arises if either the policy costs are large, or if substitution between capital and labor is very elastic. This intuition is first illustrated in a simple model where I derive simple closed-form formulas. I use these formulas, combined with sufficient statistics, to produce a first set of estimates for the elasticity. I then extend this baseline model in two directions, in order to recognize that several frictions interfere with firm's input decisions and that, therefore, managers have an incentive to evade the regulations. I develop a more comprehensive, structural model that incorporates two extra margins: a labor friction that rationalizes the presence of firms in the dominated region after the policy thresholds, and the possibility for managers to misreport the true number of employees. This structural approach delivers similarly low firm-level elasticities.

These two additional margins are considered in order to address two common concerns with bunching designs, which can potentially bias my estimates. First, the economic effect of the policy can be *overestimated* if the observed response is mainly due to evasion or misreporting.[4] Second, the structural elasticity can be *underestimated* if the agents fail to

[4]See Henrik Jacobsen Kleven 2016: "in contexts in which evasion and avoidance responses are feasible, observed bunching can be large in elasticity terms" and also Saez 2010, Chetty, Friedman, et al. 2011, Henrik J Kleven and Waseem 2013, Bastani & Selin (2014).

notice or understand the policy, or if their decisions are dampened by frictions.[5] I take these two concerns seriously and build a framework designed to tackle both evasion and salience. To detect evasion, I confront two administrative datasets and use the gap between the alternative measures of employment at the firm level to detect evasion.[6] The first measure is self-reported in the appendix of the firms' tax returns. The firm size distribution calculated from these measures shows very visible bunching patterns before the policy thresholds, which rules out the concern that these policies fall unnoticed by the firm managers.[7] I then merge this first dataset with payroll data that is extensively checked by the social security administration. The bunching on this second dataset is much more muted, suggesting that misreporting might be an important issue. Instead of focusing on just one of these measures, my estimation procedure uses the information contained in the joint distribution. In particular, the distribution of the gap between these two measures is informative of the *perceived* cost of the policies on which the firm's manager based her hiring and investment decisions.

In the data, misreporting is concentrated among the firms located before the threshold. Evasion drives a large part of the bunching response. The bunching mass, once evasion is properly taken into account, is 20% to 25% smaller than the naive bunching mass based on the self-reported counts. The firm-level elasticities I obtain lie around 0.1 and are on the lower spectrum of the estimates in the literature. The elasticities are the highest in Business Services and Wholesale Trade, sectors who tend to be less capital intensive. I repeat this exercise at the various thresholds and in the main sectors of the economy. I find very little evidence of capital labor substitution, either when looking directly at distortions in capital labor ratios or by inspecting investment behavior in the threshold in a dynamic extension of my baseline framework.The conclusion of this exercise is that micro-level production func-

[5]See Saez 2010, Henrik J Kleven and Waseem 2013

[6]Garicano, Lelarge, and Van Reenen 2016 used the second source as a robustness check but they do not examine the joint distribution of the labor measures.

[7]There is no shortage of evidence that firm managers feel strongly about these size-based regulations. For at least half a century (see Gattaz 1979), policy recommendations about these regulations emanating from French major trade groups or their representatives read like a litany of calls for drastic simplification or outright suppression. A law voted in Fall 2018 in the French National Assembly trimmed some of the regulations around the 20-employee threshold, a partial granting of these demands that fell short of a more thorough simplification.

tions are quite rigid and look more like Leontief production functions. As a consequence, flexibility at the aggregate level and in the longer-run must arise from reallocation between firms or from the entry of firms with radically different technologies. To assess the potential magnitude of these channels, I build up the aggregate production function from the micro production functions as in Baqaee and Farhi 2018 and Oberfield and Raval 2014. The extent of the reallocation across firms is empirically limited by the amount of heterogeneity within sectors. I find that given the amount of heterogeneity measured in the data, the aggregate elasticity of the economy is about 0.3. This is lower than the consensus of recent estimate, but similar to the values found in Lucas 1969. As a consequence, an active extensive margin, i.e. the entry and exit of firms, is important in order to obtain the elasticity of substitution in the range commonly assumed.

There are several advantages to my approach. First, the quasi natural experiment setting is not vulnerable to the usual endogeneity issues and bias towards one that characterize the approaches based on aggregate time-series or cross-sectional variations. Second, the size of the French economy is large enough so that I can estimate the elasticities for 2-digit sectors, across which technologies and substitution can in principle be different. Another advantage of my setting is to be able to estimate the elasticity at several points in the size distribution of firms, corresponding to the different regulatory thresholds. The 50-employee regulatory threshold is the most visible in the firm size distribution and the most notorious. I also consider the 10- and 20-employee thresholds. While there is hardly any reaction around the 20-employee threshold, firms do respond to the 10-employee threshold, and I use the larger number of firms in this range of employment to sharpen the power of my estimates. Third, my approach is flexible and can be adapted to other contexts of size-dependent policies since I derive closed-form formulas for the elasticity of substitution under a variety of wedge specifications. This is particularly useful in some contexts where the regulatory cost comes from a collection of policies, contrary to the usual implementation of the bunching estimator in the public finance literature where the precise cost of the policy is known.

This work builds on a number of contributions. While growth theory has conventionally relied on Cobb-Douglas production functions, which have a unitary elasticity (Solow 1957),

most empirical studies reject this knife-edge assumption (see Chirinko 2008 for an extensive survey) and conclude that the range $0.4-0.6$ is a reasonable ballpark for this elasticity. These studies rely on either a combination of strong structural assumptions coupled with time-series variations or try to make use of plausibly exogenous cross-sectional variations in factor prices. Oberfield and Raval 2014 follow the second approach and obtain an aggregate elasticity of 0.6 that is obtained from aggregating up micro elasticities. I depart from their work in several dimensions. First, instead of estimating plant-level elasticities from cross-sectional and spatial variations in the factor prices, I use very different identification assumptions: a quasi-natural setting and bunching techniques. My setting is therefore not subject to the endogeneity issues in Oberfield and Raval 2014 which would bias the elasticity estimate towards one. This bias is driven by the correlation in the time and spatial variations of the wages, with the location and input decisions of firms (e.g. more labor-intensive firms locating in regions with lower wages). Second, and perhaps for this reason, I found elasticities that are half as small as theirs. My finding that elasticities are close to zero is congruent with recent work in the empirical literature on production networks, such as Atalay 2017, Boehm, Dhingra, Morrow, et al. 2016, and Barrot and Sauvagnat 2016, who all find structural elasticities of substitution in production that are significantly below one. In a recent work Doraszelski and Jaumandreu 2018also structurally estimate small elasticities of substitution in an unbalanced panel of Spanish firms. This finding cast doubt on the ability of the decline in the price of equipment to serve as an explanation for the fall of the labor share, which would require an elasticity higher than one. More precisely, my results suggest that this mechanism is unlikely to occur at the *intensive* margin, that is, for firms already operating in the economy. In contrast, this mechanism could be important on the *extensive* margin of capital-labor substitution, i.e. through patterns of entry and exit, if newly created firms are able to more flexibly choose their factor ratios. Recent evidence, however, points to a decline entry and exit rates (Decker et al. 2017), which limits how much traction could be gained through this channel.

The use of bunching designs to estimate tax elasticities has blossomed in the public finance literature following the seminal work of Saez 2010 who studied kinks in taxpayers marginal

income tax rates and found that the actual response was much smaller than the one that could be expected from the theory.[8] Notches, which create a discontinuity in the decision sets, elicit larger responses.[9] A recurrent concern, though, is whether the observed responses are due to evasion or not. As emphasized by Saez 2010 for the case of kinks, bunching designs are potentially vulnerable to misreporting and other manipulation, especially in absence of third-party reporting. [10] By jointly exploiting the two datasets, I establish that evasion is behind a large part of the bunching observed. By contrast, taking the self-reported data at face value leads to an overestimation of the true elasticity, as it not only overstates the amount of bunching but also inflates the capital intensity distortion by artificially shrinking the denominator. Closest to the spirit of my paper is Chen et al. 2017's investigation of a Chinese R&D subsidy program. They use a notch design in order to uncover the returns on investment and are also confronted with potential misreporting issues, that they control for indirectly using proxy variables.

Assessing the aggregate impact of size-dependent policies has been an active research topic in macroeconomics. Restuccia and Rogerson 2008 developed one of the first framework to study the economy-wide impact of the misallocation generated by these policies. Inferring the aggregate effects of these distortions from the firm size distribution is a popular diagnostic tool in the macro-development literature (see Hsieh and Olken 2014 for a survey, or Bachas and Soto 2018 for a recent application). These aggregate implications can be either inferred from wedges taken as primitives or directly traced back to specific policies. I follow the first approach and capture the bundle of policies with wedges. While these concerns have been progressively integrated in the applied and public finance literature (see Henrik Jacobsen Kleven 2016 and references therein) they have not been fully recognized in the large macroeconomic literature on misallocation, which cites the bunching of French firms around

[8] see Henrik Jacobsen Kleven 2016for an extensive survey.

[9] See Chetty 2012: "Audit studies show that self-employment income is frequently misreported on tax returns because of the lack of double reporting. (...) Unlike kinks, notches in budget sets, where a $1 change in earnings leads to a discontinuous jump in consumption, generate substantial behavioral responses. (...) *Notches are therefore a promising source of variation for identification of structural elasticities.* "

[10] See for example Carrillo, Pomeranz, and Singhal 2017 and reference therein for a study of the effect of third-party reporting on evasion.

these regulatory thresholds as a canonical example of distortionary, size-dependent policies with potentially significant welfare-reducing effects (see Garicano, Lelarge, and Van Reenen 2016 for a welfare analysis, and Hopenhayn 2014 for a survey of the misallocation literature). Because of its prominence in the macroeconomic literature, clarifying the role of evasion in the French is important.

Finally, the 50-employee threshold in France has attracted the attention of both policy-makers and researchers. Recently, Garicano, Lelarge, and Van Reenen 2016 and François Gourio and Roys 2014 have studied the 50-employee threshold in France to estimate its welfare impact. While these two papers are concerned with the potential misallocation of factors induced by these policies, the focus of this paper is rather different: I am looking at these policies only to the extend that they provide a quasi-natural setting to estimate the elasticity of substitution. Moreover, Garicano, Lelarge, and Van Reenen 2016 and François Gourio and Roys 2014 reach divergent conclusions as to the significance of the welfare effects. In a dynamic model with sunk costs, François Gourio and Roys 2014 find the aggregate effects to be negligible while Garicano, Lelarge, and Van Reenen 2016 in their baseline specification estimate a welfare cost amounting to 3.4 of GDP. These large effects might seem at odds with most of the studies reviewed in Hsieh and Olken 2014, who express skepticism about the potency of size distortions on medium-size firms. I show that the strong bunching pattern used in their estimation could be in large part an artifact of misreporting.[11] Therefore, acknowledging the role of misreporting could help reconcile their findings with Hsieh and Olken 2014's conclusions. Unlike these studies I do not assume that the distribution of firms' productivities is Pareto, and relax assumptions on the production function in order to analyze capital labor substitution.

The rest of the paper is organized as follows. Section 2 lays down the general framework to aggregate micro elasticities into macro elasticities. Section 3 develops an heterogeneous firms model and derives closed-form formulas that link the elasticity of substitution and the

[11]This view is confirmed by ongoing work by Askenazy and Breda, who carefully investigate the institutional analysis of the misreporting phenomenon. The possibility that misreporting could be important was raised in Ceci-Renaud and Chevalier 2010.

policy cost with the bunching mass and the capital distortion patterns. Section 4 describes the institutional background and the data. Section 5 details the estimation and presents the main results. Section 6 confirms these findings by estimating a fully-fledge structural framework that incorporates friction and an additional margin for evasion, and Section 7 concludes.

1.2 The Aggregate Production Function and Firms' Micro-Elasticities

How does the substitution between capital and labor in the economy relates to elasticities of substitution at the firm level? To answer this question, I build explicit foundations for the aggregate production function in this economy using the rich framework developed by Baqaee and Farhi 2018 and a series of follow-up articles. It is important to recognize that the micro and macro elasticities can in theory be completely disconnected, a point was first raised by Houthakker 1955. In his framework, a continuum of firms using only Leontief technologies ($\sigma = 0$) aggregates into a Cobb-Douglas production function ($\sigma = 1$). This stark result was later extended by Levhari 1968. Any aggregate CES production functions can a emerge from a suitable distribution of Leontief units.

Several mechanisms contributing to making the aggregate elasticity of substitution potentially larger than the micro-elasticity: (i) substitution across production units, (ii) substitution between sectors, and (iii) entry of new units with different capital intensities. These possibilities have long been recognized.[12]. The first two channels are explicitly present in my framework, which has a nested CES structure.This framework delivers a simple expression of the aggregate elasticity, that only depends on the firm-level elasticities and sufficient statistics that capture the heterogeneity and shares of each sectors. [13] Final demand is a

[12]Hicks 1932 conjectured that in a multi-sector economy, the aggregate elasticity of substitution is greater, (a) the greater the intra-sectoral elasticity, (b) the greater the difference in factor intensity among sectors, (c) the greater the inter-commodity substitution by consumers, and (d) the greater the technological innovation that enhances intra-sectoral and inter-commodity substitution (see Miyagiwa and Papageorgiou 2007). Arrow et al. 1961 suggest that "Given systematic inter-sectoral differences in the elasticity of substitution and in income elasticities of demand, the possibility arises that the process of economic development itself might shift the over-all elasticity of substitution".

[13]Reallocation between firms occurs as long as firms are heterogeneous in a meaningful way. While remi-

CES aggregator over the sectoral goods

$$Y = \left(\sum_n \omega_n Y_n^{\frac{\eta-1}{\eta}} \right)^{\frac{\eta}{\eta-1}}$$

and the output in sector n is a CES over the output of all the firms operating in that sector

$$Y_n = \left(\sum_i \omega_{ni} Y_{ni}^{\frac{\epsilon_n-1}{\epsilon_n}} \right)^{\frac{\epsilon_n}{\epsilon_n-1}}$$

where ϵ_n is the elasticity of demand, Y_{ni} the output of firm i in sector n and D_{ni} demand weights. where η is the elasticity of demand across sectors. Under these demand assumptions, the industry level elasticity is a weighted average of the micro elasticity and the demand elasticity. The more heterogeneity there is between units, the more potential for substitution across units. I suppose that firms in this economy have access to a CES production whose elasticity is fixed at the sector level but operate at potentially different factor ratios. Each firm maximizes profit taking all prices as given and market for all goods and factors clear. Therefore the welfare theorems apply and I can use Shephard's lemma to express the aggregate elasticity using the equilibrium the labor cost ratios

$$\alpha_{ni} \equiv \frac{rK_{ni}}{rK_{ni} + wL_{ni}}$$

The exact industry elasticity is a linear combination of within and between plant capital-labor substitution:

$$\sigma_n^{ind} = (1 - \chi_n)\, \sigma_n + \chi_n \epsilon_n$$

where $\chi_n = \sum_i \frac{(\alpha_{ni} - \alpha_n)^2}{\alpha_n(1-\alpha_n)} \theta_{ni}$ is a measure of the dispersion of the factor cost ratios within sectors . Similarly the aggregate elasticity is a linear combination of the sectoral elasticities and a dispersion

$$\sigma_{agg} = (1 - \chi)\, \hat{\sigma} + \chi \eta \tag{1.1}$$

where $\hat{\sigma} = \sum_n \frac{\alpha_n(1-\alpha_n)\theta_n}{\sum_{n'} \alpha_{n'}(1-\alpha_{n'})\theta_n} \sigma_n^{ind}$ and $\chi = \sum_n \frac{(\alpha_n - \alpha)^2}{\alpha\,(1-\alpha)} \theta_n$

niscent of the reallocation mechanism at play in Houthakker 1955, Levhari 1968, and Satō 1975, reallocation in my framework do not rely on distributional assumptions about firm's production *capacities*. See Oberfield and Raval 2014 for more detailed comments regarding this important distinction.

This simple framework aggregates sectors in a symmetric fashion, regardless of the actual input-output links between these sectors. However, aggregate effects could in principle be traced down to micro-elasticities, as they propagate through the input-output structure of the economy. In my baseline framework, all the firm-level productions factors have only two factors of productions. There is therefore no room for input-output linkages as they are no intermediary goods. In the appendix, I consider the full framework developed by Baqaee and Farhi 2018 with intermediary goods. I do not observe the basket of intermediary goods materials at the firm-level but I observe spendings on materials. To understand the propagation effects I can the input-output composition and scale it by the materials at the firm-level.

1.3 Micro Elasticities of Substitution and Size-Dependent policies

This section develops a new method to relate moments in the data such as the mass of bunchers and the capital intensity distortion to the structural parameters of interest and in particular the elasticity of substitution. I start with a simple span of control model *á* la Lucas 1978 in order to derive transparently the theoretical effects of the policies in the firm's input decision. The basic intuition is that some firms will respond to the exogenous change in the implicit price of labor by staying below the threshold and over-accumulating capital. In this stark model, a mass of firms should bunch right before the threshold and there should be a valley with no firms after the threshold. This prediction is at odds with the observed firm distribution in which the bump is sprayed from 46 to 49 and there is a non-zero mass of firms with 50, 51, etc. employees. Because the discontinuity in the firm-size distribution is not as sharp as the one generated by this simple theoretical setting, Section 5 introduces frictions in the input decisions. These frictions reflect the uncertainty faced by the firm owner due to the turnover of employees – such as inability to hire or unexpected quits –, or can be thought of as "optimization errors" (Chetty 2012).

1.3.1 Baseline Model.

I start with the simplest model to study size-depend policies with only one factor – labor – and then build up the general framework to show how to decompose the bunching response along the different margins. I first show how the bunching mass relates to the structural parameters in a simple, static span of control model as in Lucas 1978 where labor is the only production factor. In this baseline model, size-dependent regulations creates what the bunching literature calls a *notch* in the profit function, that is, a change in the *average* cost of production.[14] Because the very same logic applies at each policy threshold, I write the model for a generic threshold fixed at N employees. In the empirical exercise I study the firms' behavior around each threshold, that is for $N \in \{10, 20, 50\}$.

In the simplest model, I consider firm i in sector n that uses only one input - labor - and revenue productivity, z_{ni}, enters multiplicatively. The before-tax revenues of the firms are

$$p_{in}Y_{ni} = z_{ni}^{1-\nu} \cdot L_{ni}^{\nu} \tag{1.2}$$

The rest of this section analyzes the decisions of this generic firm and therefore the subscripts can be dropped to ease the exposition. The productivity index z follows an exogenous distribution with cumulative distribution function $H(.)$. I do not place any parametric assumption on $H(.)$ and only assume that it is continuous. In practice, I estimate $H(.)$ non-parametrically by local polynomials, as explained in details in Section 4.[15]

A large amount of policies come into force at the thresholds. These policies are a combination of organizational, accounting, and human resources regulations. For instance, firms need to set up a profit-sharing scheme and subsidize a workplace committee with a budget equal to at least 0.3% of the payroll. The workers committee are also awarded paid hours for training. Only a handful are expressed in monetary terms, it is thus hard to put a precise price tag on many of these measures individually. Most of these policies scale with the size of

[14] As opposed to "kinks", that are changes in the marginal cost of production.

[15] It has been customary to assume that firms' productivities follow a Pareto distribution. While this approximation is a parsimonious description for the full range of the firm size distribution, it is in fact rejected by the data (see Appendix, Figure **??** for details). Maintaining the Pareto assumption would create significant misspecification bias in our estimate of the elasticity.

the firms. Appendix B contains a detailed list of all these policies. Therefore, I capture the *combined* burden created by this bundle of regulations thus be captured as a hassle cost $\Delta\tau$ that acts as an increase in the baseline tax rate and reduces the net profit of the manager. A similar approach can be found in Aghion et al. 2017 who study in details tax regime decisions of small French entrepreneurs.[16]

For a policy threshold standing at N employees, the profit function of the firm is defined piecewise as follows:

$$\Pi(L;z) = \begin{cases} (1-\tau)\left[z^{1-\nu}L^{\nu} - wL\right] & \text{if } L < N \\ (1-\tau+\Delta\tau)\left[z^{1-\nu}L^{\nu} - wL\right] & \text{if } L \geq N \end{cases} \tag{1.3}$$

The profit function takes the usual form before the threshold and additional costs due to the regulations decrease the profits above the employee threshold. Even in such a simple setup, the consequences on the firms decisions are not completely straightforward. While the bunching literature distinguishes between two types of bunching designs, kinks and notches (Henrik Jacobsen Kleven 2016), our case does not fall neatly into either category: while the labor wedge on its own creates a kink, i.e. a change in the *marginal* cost function, the fixed cost creates a notch – a change in the average cost. From the first order condition, it is immediate that the labor decision is defined piecewise. The labor demand schedule is

$$L(z) = \begin{cases} z\left(\frac{\nu}{w}\right)^{\frac{1}{1-\nu}} & \text{if } z \leq \underline{z_0} \\ N & \text{if } z \in \left[\underline{z_0}, \bar{z_0}\right] \\ z\left(\frac{\nu}{w}\right)^{\frac{1}{1-\nu}} & \text{if } z \geq \bar{z_0} \end{cases} \tag{1.4}$$

where $\underline{z_0} = N.\left(\frac{w}{\nu}\right)^{\frac{1}{1-\nu}}$ is the productivity index of the theoretical lowest buncher. Note that labor is simply a linear function of the unobserved productivity. The productivity index of the theoretical top buncher $\bar{z_0}$, is defined by a profit indifference condition. For this marginal firm, the profit is the same whether the firm's manager decides to bunch at N or to pay the

[16]In contrast, Garicano, Lelarge, and Van Reenen 2016 choose to combine a labor wedge and a fixed cost and find that this fixed cost to be negative. François Gourio and Roys 2014 choose a sunk cost in a dynamic model.

regulatory cost and pick the unconstrained optimal employment above the threshold $L(\bar{z_0})$. To see this, notice that the equilibrium profit function is

$$\Pi(z) = \begin{cases} (1-\tau)(1-\nu) z \left(\frac{\nu}{w}\right)^{\frac{\nu}{1-\nu}} & \text{if } z \leq \underline{z_0} \\ (1-\tau)\left[z^{1-\nu}N^{\nu} - wN\right] & \text{if } z \in \left[\underline{z_0}, \bar{z_0}\right] \\ (1-\tau-\Delta\tau)(1-\nu) z \left(\frac{\nu}{w}\right)^{\frac{\nu}{1-\nu}} & \text{if } z \geq \bar{z_0} \end{cases} \tag{1.5}$$

Hence, the profit indifference condition for the top marginal buncher with productivity index z_0 is as follows

$$\Pi\left(L\left(\bar{z_0}\right); \bar{z_0}\right) = \Pi\left(N; \bar{z_0}\right)$$

Using the expression for profits, this indifference condition is

$$(1-\tau-\Delta\tau)\,\bar{z_0}\left(\frac{\nu}{w}\right)^{\frac{\nu}{1-\nu}} = (1-\tau)\left[\bar{z_0}^{1-\nu}N^{\nu} - wN\right] \tag{1.6}$$

which, after substituting out the threshold $N = \underline{z_0}\left(\frac{\nu}{w}\right)^{\frac{1}{1-\nu}}$ and rearranging, yields

$$\frac{\bar{z_0}}{\underline{z_0}} = \left[\frac{1-\tau}{1-\tau-\Delta\tau}\right]\left(\frac{1}{1-\nu}\right)\left[\left(\frac{\bar{z_0}}{\underline{z_0}}\right)^{1-\nu} - \nu\right] \tag{1.7}$$

Because of the notch in the profit function, the marginal top buncher hires $\bar{N} = L(\bar{z_0}) > N$ workers and leaves a gap in the firm size distribution. It is important to realize that the size of the theoretical hole is independent of any assumption on the unobserved productivity distribution. It is therefore useful to express equation (1.7) solely in terms of the size of the theoretical gap in the distribution with the cost of the regulation. In keeping with the bunching literature I write $\bar{N} = N + \Delta N$, where ΔN is the size of the hole after the threshold.

Proposition 1. *The size of the gap is strictly increasing with the policy cost and we have*

$$\frac{\Delta\tau}{1-\tau} = 1 - \left(\frac{1}{1-\nu}\right)\left[\left(1+\frac{\Delta N}{N}\right)^{-\nu} - \nu\left(1+\frac{\Delta N}{N}\right)^{-1}\right]. \tag{1.8}$$

Because N is known and $\bar{N}$ is in principle observable, the hassle cost can be inferred from this simple formula given a value for the returns to scale ν.[17]If the regulations creates no

[17]This formula is thus the counterpart to equation 5 in Henrik J Kleven and Waseem 2013.

distortion at all, i.e. if $\Delta\tau = 0$ then there is no bunching, $\Delta N = 0$ and $\bar{z_0} = \underline{z_0}$. Finally, the labor schedule 1.4, together with the productivity distribution $H(.)$, defines the firm size distribution $S_0(.)$. If the unobserved productivity follows a distribution with $H(.)$, then the firm size density is given by

$$S_0(n) = \begin{cases} \left(\frac{w}{\nu}\right)^{\frac{1}{1-\nu}} h\left(n\left(\frac{w}{\nu}\right)^{\frac{1}{1-\nu}}\right) & \text{if } n \leq N-2 \\ H(\bar{z_0}) - H\left(\underline{z_0}\right) & n = N-1 \\ 0 & N \leq n \leq \bar{N} \\ \left(\frac{w}{\nu}\right)^{\frac{1}{1-\nu}} h\left(n\left(\frac{w}{\nu}\right)^{\frac{1}{1-\nu}}\right) & \text{if } n \geq \bar{N} \end{cases} \tag{1.9}$$

In practice, I estimate h non-parametrically from the observed firm size distribution. As mentioned previously, this basic model with a notch has stark predictions: a mass point just before the threshold and a hole with no firms at the right of the threshold. This predicted hole, as in many bunching designs with notches (e.g. Henrik J Kleven and Waseem 2013), is clearly counterfactual as can be seen from Figure . In Section 2.3 I show how the mass of firms in that region can be understood as the consequence of optimization frictions. Before that, I analyze how capital/labor substitution in the production function increases the amount of bunching and can be measured from distortions in input choices.[18]

1.3.2 Capital Labor Substitution.

Now suppose that the production function F allows for capital-labor substitution. I show in this section that the amount of firms bunching at the threshold and the structural parameters are linked by a transparent formula. The revenue production function of a firm is given by

$$pY = z^{1-\nu} \cdot F(K, L)^{\nu} \tag{1.10}$$

[18]Assuming that the underlying productivity follows a Pareto distribution leads to counterfactual results and substantial misspecification errors. If $H(.)$, then $h(x) \propto x^{-\eta-1}$ and $\ln h(x) \propto \ln x$. It follows that the logarithm of the cdf, $\ln S_0(n)$, would be proportional to $\ln n$ with the same slopes on both sides of the threshold. This prediction is rejected in the data (see Appendix Figure **??**). This observation leads us to reject the usual Pareto assumption to a more flexible approach that measure the bunching mass more accurately.

where the productivity index z follows an exogenously given distribution, with a cdf $H(.)$ that will be estimated non-parametrically. The elasticity of substitution between capital and labor is defined by

$$\sigma(K,L) \equiv -\frac{d\ln K/L}{d\ln F_K(K,L)/F_L(K,L)} \tag{1.11}$$

With only two production factors, it corresponds to the inverse of the slope of the isoquant of the production function. When the input markets are perfectly competitive, the factors of production are paid their marginal costs and the elasticity simply relates the relative demand of factors to their relative prices, that is

$$\sigma(K,L) = -\frac{d\ln {}^{K}/_{L}}{d\ln {}^{r}/_{w}} \tag{1.12}$$

In this article I focus on the case where firms belonging to the same sector have access to the same production function F. I suppose that F is defined by the following CES production function [19].

$$F(K,L) = \left[\alpha K^{\frac{\sigma-1}{\sigma}} + (1-\alpha) L^{\frac{\sigma-1}{\sigma}}\right]^{\frac{\sigma}{\sigma-1}}$$

with $\alpha \in (0,1)$. In order to simplify the exposition, I have assumed here that firms belonging to the same sector operate with the same capital-labor ratio. When I take the model to the data, I let αbe a function of $z$such that the capital intensity in my model increases with size as in the data. The CES production function nests three special cases. The familiar Leontieff production obtains when $\sigma \to 0$. In this case there is no substitution possibilities between capital and labor and the production function writes $F(K,L) = \min\{\alpha K, (1-\alpha)L\}$. On the other hand, taking the limit $\sigma \to 1$ yields the Cobb Douglas production function $F(K,L) = K^{\alpha}L^{1-\alpha}$. Finally as $\sigma \to \infty$ the production function becomes linear and capital and labor are perfect substitutes.

Except at $L = N$, where the firms are bunching, the solution to the firm's maximization problem is interior and the first order conditions are satisfied. It follows that the capital-labor

[19] More generally, I show in the appendix that the first order conditions provides a nonparametric local identification of F.

ratios implied by this model are constant and lie in an interval at $L = N$:

$$\begin{cases} \frac{K}{L} = k^* \equiv \left(\frac{\alpha}{1-\alpha}\frac{w}{r}\right)^{\sigma} & \text{if } L \neq N \\ \frac{K}{L} = [k^*, k^* + \Delta k] & \text{if } L = N \end{cases} \tag{1.13}$$

To simplify notations, let $f(k)$ denote the intensive-form production function, that is, $f(k) \equiv F(k, 1)$. For all the firms that are not bunching at the threshold, the labor demand is linear in the productivity index

$$L(z) = z \cdot L_1 \tag{1.14}$$

where $L_1 = \left(\frac{\nu f(k^*)^{\nu}}{rk^* + w}\right)^{\frac{1}{1-\nu}}$. In the bunching region, where $L = N$,[20] the first order condition in capital is satisfied and yields

$$k(z)^{\frac{1}{\sigma}} \left[\alpha k(z)^{\frac{\sigma-1}{\sigma}} + 1 - \alpha\right]^{\frac{(1-\nu)\sigma-1}{\sigma-1}} = z^{1-\nu}\frac{r\nu N^{-\nu}}{\alpha} \tag{1.15}$$

This equation implicitly defines the equilibrium capital intensity, $k(z)$, as a function that is strictly increasing in z.[21]Intuitively, with an elasticity close to 0, the shape of $k(z)$ is flat as the production function exhibit almost little substitution whereas in the Cobb Douglas case, the relationship is exact and we have $d\log k = \frac{1-\nu}{1-\alpha\nu} d\log z$. The relative productivity gap between the top and bottom marginal bunchers is $\frac{\Delta z}{z} = \frac{\Delta N}{N}$. Therefore, if the hole in the distribution has size 10 ($\frac{\Delta N}{N} = 0.2$), a distortion in the capital ratio of about 6% is expected in the Cobb-Douglas case. For the more general CES case I solve this equation numerically and provide a first order approximation.

Proposition 2. *The amount of distortion in the capital intensity for the firms in the bunching region increases with the elasticity of substitution if $\sigma < 1$. At the first order approximation, we have*

$$\frac{\Delta k}{k} \approx \sigma \frac{1-\nu}{1-\alpha\nu}\frac{\Delta z}{z}$$

This relationship captures the link between the unobserved productivity and the distortion. However, the productivity is not an object that is observed. Instead, the capital distortions in terms of observable revenues.

[20]More precisely, it is equal to $N - 1$ since this value is reported as an integer. I abstract from the integer problem in the stylized model but return to this issue in more details in the empirical section.

[21]This monotonicity holds in general for a larger class of productions F and is not specific to the CES.

The profit functions are also linear in the productivity index, except in the bunching region (see Figure **??**), we have

$$\begin{cases} \Pi^*(z) = (1-\tau)\, z \cdot \Pi & z < \underline{z_1} \\ \Pi^*(z) = (1-\tau)\left[z^{1-\nu} N^{\nu} f(k(z))^{\nu} - rk(z)\, N - wN\right] & \text{if } z \in \left[\underline{z_1}, \bar{z_1}\right] \\ \Pi^*(z) = (1-\tau-\Delta\tau)\, z \cdot \Pi & z > \bar{z_1} \end{cases} \tag{1.16}$$

where $\Pi \equiv \left(\frac{\nu f(k^*)^{\nu}}{rk^*+w}\right)^{\frac{\nu}{1-\nu}}$ is a constant that corresponds to the before-tax profit of a firm with productivity index $z = 1$. The index of the lowest buncher, $\underline{z_1}$, is determined by $L\left(\underline{z_1}\right) = \underline{z_1}\left(\frac{\nu f(k^*)^{\nu}}{rk^*+w}\right)^{\frac{1}{1-\nu}} = N$ and the unobserved productivity index $\bar{z}_1$ of the top buncher is determined by the following indifference condition

$$(1-\tau-\Delta\tau)\,\bar{z}_1 \cdot \Pi = (1-\tau)\left[\bar{z}_1^{1-\nu} N^{\nu} f(k(\bar{z}_1))^{\nu} - rk(\bar{z}_1)\, N - wN\right] \tag{1.17}$$

To obtain a formula similar to the one derived in the previous section, I take a first order approximation of equation1.17 . The capital intensity is increasing in the bunching region and I write $k(\bar{z_1}) = k^* + \Delta k$. Using the first order condition in capital taken at $z = \underline{z_1}$ and the fact that, in equilibrium, the before-tax profit is equal to a fraction $1-\nu$ of output, we have

$$\frac{\bar{z_1}}{\underline{z_1}} = \frac{1-\tau}{1-\tau-\Delta\tau}\frac{1}{1-\nu}\left[\left(\frac{\bar{z_1}}{\underline{z_1}}\right)^{1-\nu} - \nu + \frac{\Delta k}{k^*} s_K \left[\left(\frac{\bar{z_1}}{\underline{z_1}}\right)^{1-\nu} - 1\right]\right] \tag{1.18}$$

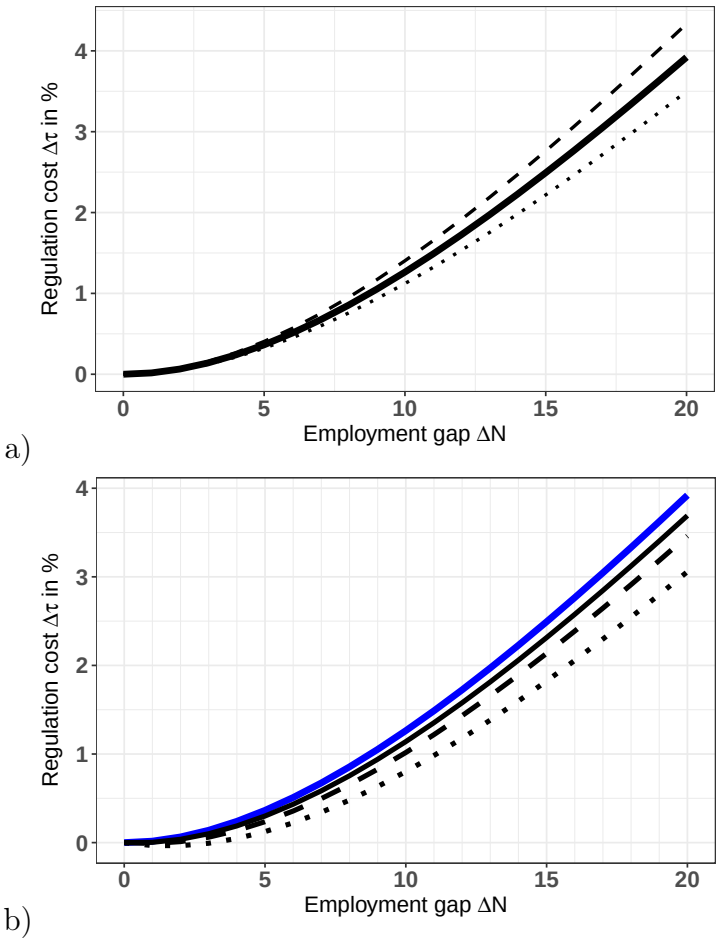
where s_K is the revenue share of capital. The bunching formula (1.18) is the counterpart of (1.7) when the manager can substitute between capital and labor. Firms in the bunching region have an incentive to boost their capital intensity instead of crossing the threshold. The extra profits generated by this increase in capital intensity is captured by the last term in the brackets. After substitution, this formula yields a relationship between the size of the hole after the threshold, the policy hassle cost and and the distortion.

Proposition 3. *The amount of bunching at the threshold increases with the policy cost and with the elasticity of substitution, and the three are related by the following equation*

$$\frac{\Delta\tau}{1-\tau} = 1 - \left(\frac{1}{1-\nu}\right)\left(\left(1+\frac{\Delta N}{N}\right)^{-\nu} - \nu\left(1+\frac{\Delta N}{N}\right)^{-1} + \frac{\Delta k}{k^*} s_K \left[\left(1+\frac{\Delta N}{N}\right)^{-\nu} - \left(1+\frac{\Delta N}{N}\right)^{-1}\right]\right)$$

where s_K is the capital share.

Figure 1.1: Relationship between elasticity, regulation costs, and the theoretical employment gap.



Panel a) plots the relationship between the regulation cost and the size of the hole in the firm size distribution in the baseline one-factor model. Panel b) represents the formula computed for $\sigma = 0$ (solid), $\sigma = 0.1$ (dash), and $\sigma = 0.3$ (dotted).

This formula generalizes the formula in Proposition 2 to the case where capital and labor are substitutable. In the extreme case where there is no substitution at all, then the capital intensity stays the same in the bunching region and $\frac{\Delta k}{k^*} = 0$. In this case the formula reduces to the perfectly rigid case covered by Proposition 2. This formula maps three observables, the labor share, the capital intensity distortion and the gap in the size distribution ΔN into the policy cost. Different combinations of policy costs and elasticities can account for a given level of excess mass. Figure (1.1) plots the combinations of elasticities and policy costs that

can jointly generate the observed excess mass.

The firm size density, $s_1(.)$ is a simple linear transformation of the productivity density $h(.)$, which directly follows from the labor demand 1.4. It is defined piecewise as follows

$$s_1(n) = \begin{cases} \frac{1}{L_0} h\left(\frac{n}{L_0}\right) & \text{if } n \leq N-2 \\ H\left(\frac{\bar{N}}{L_0}\right) - H\left(\frac{N}{L_0}\right) & n = N-1 \\ 0 & N \leq n < \bar{N} \\ \frac{1}{L_0} h\left(\frac{n}{L_0}\right) & \text{if } z \geq \bar{z}_1 \end{cases}$$

That the firm-size density adopts such a simple expression facilitates the empirical work, as a simple rescaling delivers the unobserved productivity distribution from the firm size distribution. This allows me recover the unobserved productivity distribution $h()$ non-parametrically instead of having to assume a strong functional form. Relaxing these functional forms is important as it removes the dependency of the policy cost and elasticity estimates on the unobserved productivity distribution. This problem is even more acute if one views these regulations as a combination of a fixed cost and a variable cost. In this case, these policy costs are not separately identified and the identification ultimately rests on the choice of functional for for the unobserved productivity (see appendix B for a detailed proof).

The (gross) welfare cost of these policies in this model is transparent can be broken down into two pieces. The first part simply corresponds to the hassle cost $\Delta\tau$ multiplied by the before-tax profits. Given the decreasing returns, this cost corresponds to a fraction $1-\nu$ of output produced. The second component of the welfare cost is due to the local distortions at the threshold. By refusing to grow and distorting their capital intensity, firms uses resources inefficiently. This welfare cost can be viewed as an upper-bound on the effective welfare cost of these regulations since, in the model, the cost of the policies is captured exclusively by a hassle cost for the manager. To the extent that some of the regulations achieve their intended goal of providing workers with more safety (the compulsory safety committee), less job uncertainty (firing rules), or better incentives (the profit-sharing schemes), the *net* welfare cost might be smaller. Another concern is that parts of the payments made by the

firm are compensated by a decrease in the wages. However, decreasing the employees' wages to implement a profit-sharing scheme is illegal[22] so this cancelling effect could only happen slowly overtime in terms of dampened wage growths.

1.4 Data and Institutional Background

This section presents the data and the institutional background. I test the theory using two comprehensive administrative datasets. Merging these two datasets allows me to use several measures of employment and look for direct evidence of evasion. Evasion is by nature hard to measure, by using the two datasets jointly I can capture evasion behavior better precision than with the traditional approach of using some proxies.[23] Both datasets are generated by the administration and cover the universe of all French firms that have at least one employee . We merge these two datasets using the unique firm and establishment identifiers created by the administration. The first dataset is generated from the tax retunrs while the second one comes from payroll data. Since these datasets are not new to the macro/labor literature I focus only on the aspects that are the most relevant for this paper and refer to Bagger et al. 2014 for instance for details.

1.4.1 Regulations.

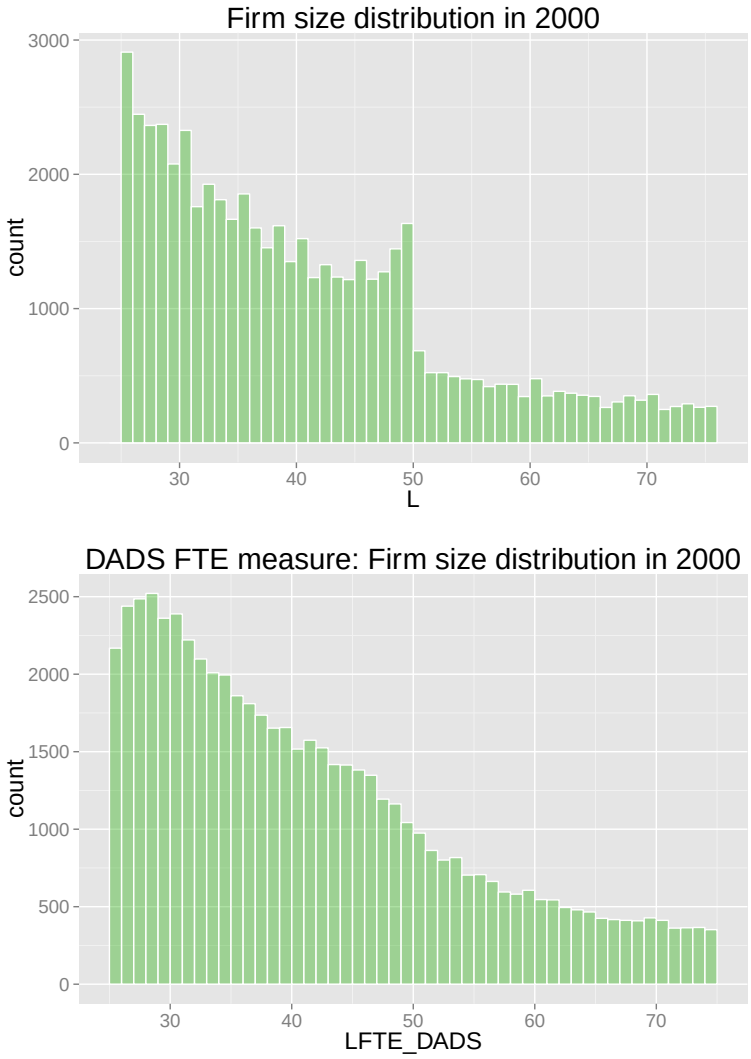
A complete list of the regulations can be found in Ceci-Renaud and Chevalier 2010. The most important policies that are activated at the 50-employee threshold are: (i) an elected workplace committee must meet once every other month. Elections must be held (unless there is no candidate); (ii) information must be transmitted quarterly and yearly about the firm's financial situation to the workplace committee; (iii) external "representative" trade unions can designated a union representative; (iv) a workplace safety committee must be

[22]Ordonnance n° 86-1134 du 21 octobre 1986 relative à l'intéressement et à la participation des salariés aux résultats de l'entreprise et à l'actionnariat des salariés.

[23]For instance, Chen et al. 2017, Bachas and Soto 2018, or Hurst, G. Li, and Pugsley 2014 look at distortions or breaks in variables potentially related to their main variable interest.

elected, with up to 3 members, which 2 paid hours every month to each member in order to run the committee; (v) compulsory annual negotiation about wages, work duration, etc; (vi) if the firm increases dividends with respect to the previous two years then a bonus must also be paid to employees; (vii) a specific procedure ("Plan de Sauvegarde de l'Emploi") must be put in place a mass layoff is considered (where mass layoff is defined as the layoff of at least 10 employees over less than 30 days); (viii) workplace negotiations must be held about prevention of accident injury and work difficulty; (ix) a compulsory monthly reporting of job turnover must be sent to the administration

Figure 1.2: Firm Size Distribution around 50-employees in the two administrative datasets



The top panel shows the firm size distribution using the employment reported on the tax form. The bottom shows the firm size distribution using the measure checked by the administration. Sources: FICUS and DADS 2000

1.4.2 Tax return dataset.

The first dataset is created by the Tax administration from mandatory corporate tax returns. The tax form[24] contains the detailed balance-sheet and income statement information: assets,

[24]The official form is CERFA 2050-9 . It is the equivalent of the IRS Form 1120 for Corporate Income Tax Return.

broken into different types, as well as investment, liabilities, wage bill and materials. Some of these variables are retreated by the administration to reduce errors. The methodologies used to construct and cross-validate some of the variables have evolved over time and this is reflected in several versions of the datasets. I follow Garicano, Lelarge, and Van Reenen 2016 and use FICUS, which is the most comprehensive one. I observe several measures of capital, gross and net values. The net value correspond to the gross value minus the accumulated depreciation, which is also observed. Capital is broken down into several categories and it is in principle possible to look separately at tangible and intangible capital. Intangible capital represent on average 20% of the gross non financial assets. It is notoriously difficult to measure capital precisely (see Becker et al. 2006 for an extensive discussion of the numerous challenges). The book value, which is most commonly available in micro data, can differ substantially from the economic value, the relevant concept for the production function. Available measures of capital therefore tend to exhibit substantial measurement error[25] and differ from the concept of capital used in production function. I use alternatively several measures. First, I use the net value of plant, property, and equipment capital, the closest counterpart in the French data of the variable called PPENT in Compustat[26], used in most U.S. studies. For consistency reasons, I focus on the period 2000-2007. My panel is not long enough to make it reliable to use a perpetual inventory method to construct a measure of capital. For some of the robustness checks, I use different measures of investment that distinguish different kinds of tangible and intangible investments. Some of these variables are only available in the BRN dataset, also generated by the administration from the tax returns, but with a coverage that excludes self-employed companies that report under the simplified tax regime.

The FICUS dataset also contains the number of employee. More precisely, firms are asked to report the arithmetic average of the employees headcount at the end of each quarter. Importantly, this number is reported in the appendix of the corporate income tax form

[25]See Collard-Wexler and De Loecker 2016: "There is reason to believe that, more than any other input in the production process, there are severe errors in the recording of capital stock." See also Kim, Petrin, and Song 2016 for a two-step estimation method in presence of measurement error.

[26]PPENT stands for Property, Plant, and Equipment Net Total.

and, importantly, is *not* checked at all by the tax administration, which indeed warns of the "mediocre quality" of this measure. [27]This can partly be blamed on the official form[28] containing the explaination on how to fill up the tax return.

1.4.3 Social Security dataset.

The second main dataset I use is a matched employer-employee dataset generated from Social Security data, called DADS (Donnees Annuelles des Declarations Sociales). Since it has been used several times in the macro labor, see for instance Bagger et al. 2014, I will only emphasize the points that are most important for the purpose of this paper. The data are based upon mandatory employer reports of the gross earnings of each employee subject to French payroll taxes. These taxes apply to all "declared" employees and to all self-employed persons, essentially all employed persons in the economy. The data comes from mandatory reports by firms to the administration that are used to compute various social security contributions as well as rights to unemployment benefits, health insurance, work accident benefits and retirement. It contains information about each job spells, the number of hours worked, the type of job, the net and gross salary etc. These can be aggregated at the establishment or the firm levels. The social security administration computes three employment measures at the firm level: 1) a Full-Time Equivalent of the total number of employees par firms 2) a headcount on December 31, 3) an arithmetic average of the headcount at the end of each quarter. As detailed below, the concept that is the closest to the one used in the regulations is the Full-Time Equivalent count. Notice that it is also the one that is the closest in spirit to the one that should enter the production function.

[27]"Il existe également des informations sur l'emploi dans les liasses fiscales, mais de médiocre qualité (beaucoup d'informations manquantes en particulier)" Beguin and Haag 2017, p.20.

[28]Form N° 2032-NOT-SD specifies how to fill in forms n° 2050 to 2059-G that constitute the corporate income tax returns for firms under the general regime.

1.4.4 Legal workforce concepts.

In the French Labor Code, the relevant concept for the workforce is the Full-Time Equivalent employment. Almost all the policies coming from the Labor Code hinge on the FTE measure. There can be, however, a slight ambiguity as to whether this FTE count must be made at the firm of at establishment level. Article L1111-2 of the French Labor Code lays down the definition of the workforce to be used for all the Labor Code regulations: (i) full-time employees with open-ended contracts count for one unit; (ii) employees on a fixed-duration contract, on intermittent contract, employees "leased" by another company that work in the company premises since at least a year, and temporary workers are to be counted in proportion of their presence over the last twelve months; (iii) part-time employees, regardless of their working contract, are counted in proportion of the number of hours written in their contracts divided by the legal or conventional work time duration. There is one exception: employees leased by another company who replace an employee on maternal/paternal leave are not counted. The regulations contained in the other Codes mostly follow the same definition.

1.4.5 Do firms report labor accurately?

Evasion is a serious concern in many bunching designs, since evasion that can substantially bias the estimates. Direct evasion is by nature often hard to observe in administrative datasets. An advantage of my setting is that by combining two measures of employment coming from different administrative source, I am able to provide direct evidence that firms under-report their workforce around the thresholds. This evidence suggests that evasion drives the majority of the observable response. To make this point more rigorously I explore in details the discrepancies between the two labor measures.

I start by looking for traces of misreporting by comparing the two firm size distributions that can be computed from the two sources. If One reason to proceed that way is that the coverage of the datasets do not perfectly coincide. In this preliminary steps I do not throw out any firms and instead analyze the relative changes in the shape of the distribution.

The first observation is that the two distributions look roughly similar. However they differ significantly by the amount of distortion around the threshold. To see this more clearly, I compute the log-ratios of the density based on self-reported employment divided by the density based on the measured FTE employment (Figure). A clear and characteristic spike appears right before the threshold. The density of firms reporting 49 employees is about 1.9 times the density of firms that have 49-employees according to the administration. The number of firms declaring 9 employees is 1.45 times larger. There is no small discernible bump around the 20 threshold but it is not significant. This is exactly the pattern we would expect if the incentives to misreports were large around the 50 and 10 threshold and weak around the 20 threshold. I repeat the same exercise in each 2-digit sector. Figure 1.9 shows the log ratios of the density for manufacturing , retail, wholesale and transportation and communication sectors . The same characteristic pattern is visible there.

Next I focus on the individual discrepancies by merging the two datasets using unique the firm identifiers and compute the difference between the two measures. I group firms by the amount of labor they report. Figure (1.10) reports the mean and the standard deviation of the gap. The significant bump in the gap just before the threshold is a clear evidence of misreporting. The gap between the two measures of labor could also be due to differences in the concept of workforce. On the tax form, firms are instructed to compute the arithmetic mean of the workforce at the end of each quarter, while on the other dataset, the Full Time Equivalent is computed. This is the relevant measure for the majorities o the policies according to the Labor Code.

There are several reasons, both in theory and in practice why the two measures do could not coincide. Counting ETP is not entirely straightforward. The methodology that the INSEE itself uses consists in several steps. Indeed, despite following their method carefully I am still left we some residual discrepancy. I thus use directly their measure. As mentioned above, the fiscal dataset contains an employee count. This count is the arithmetic average of the employees headcount at the end of each quarter. This number is reported in the appendix of the corporate tax form, and is not checked thoroughly by the tax administration.

There are also several reasons why the DADS measure is more reliable. First, while the

FICUS headcount are not verified by the administration (Beguin and Haag 2017), the DADS measures are the subject of detailed verifications. The reason for it is that employee benefits and retirement claims are based on the information submitted by the firms. The original information is generally generated from the firm's accounting and payroll software. It seems difficult to under-declare the number of workers without harming directly one or several workers individually. On the contrary, nothing really prevents firms from misreporting there employment count on their corporate tax form. After receiving the raw data, the statistical agency performs several rounds of checks. This is therefore the highest quality measure of employment according to the French National Statistical Agency.[29]

Also note that the employment count reported in the fiscal source is rounded at the nearest integer. In contrast, using the social security data and aggregating at the firm level, I obtain non-integer numbers. I compute the gap before rounding but to in order to compute the firm size distributions I round these employment counts at the nearest integer. I do not measure significant rounding at multiples of 5 or 10. Figure 1.10 shows the gap between the two measures for each firm size. The self-reported number tends to be always larger than the Full-Time equivalent measure in the DADS.

1.4.6 Primary and secondary jobs.

There is no such thing as a perfect measure of employment. In the social security data, the statistical agency computes for each worker a full-time equivalent . But this count excludes those job spells deemed "secondary" ("emplois annexes" in French). The dataset documentation contains the precise definition of these secondary jobs as welle as detailed explanations of the implementation procedure. Excluding secondary jobs is actually innocuous. While they account for about a fifth of the *observations* over the period they represent a negligible

[29]Beguin and Haag 2017 strongly recommend the use of the DADS employment count instead of one in the fiscal data: "La source administrative la plus naturelle (car à la fois la plus exhaustive et la plus fiable) pour estimer les données d'emploi dans Ésane était donc la source DADS-U2" p.150. See also Ceci-Renaud and Chevalier 2010.

share of *hours* worked.[30] A potential concern would be that these secondary jobs could be heavily concentrated in the firms bunching below the threshold. If that were the case then we would need to accommodate for a channel through which "secondary jobs" could substitute for "primary jobs", which could create a downward bias on our estimate.[31]The data shows that this is not the case. In Figure (1.13), I add up all the hours worked in secondary jobs and express them as a fraction of total hours worked. This fraction is stable around 1% and does not vary around the thresholds. Moreover, they are heavily concentrated in some specific activities among the service sectors such as: cleaning, restaurants, hotels, and various recreational and sports activities, associations.[32] In addition, some of these spells are created artificially either by the employer to register variable salary components such as bonuses.[33] The more important point is that the employment measure obtained by summing up full-time equivalent spells at the firm-level slightly underestimate the true employment. This is because the construction of FTE measure makes it top-censored at 1. [34] This underestimation

1.4.7 Compliance and Enforcement.

Are these policies really binding? Do we have evidence that firms are compliant? Given the large number of policies, it is difficult to give a definitive answer. The enforcement of

[30]In 2002 the share of secondary jobs was 24.4% but these jobs represent only 1.2% of the aggregate share of hours worked. Moreover, this share is constant across firms sizes, see appendix and Beguin and Haag 2017 p.154.

[31]Garicano, Lelarge, and Van Reenen 2016 use that figure to argue in favor of using the employment variable in FICUS and not the one in the DADS : "unfortunately, full-time equivalent information in DADS misses almost one-quarter of jobs" p.3455.

[32]Temporary workers count towards the legal employment threshold, in proportion of their time spent in the company. However they are counted neither in the fiscal dataset or in the social security dataset. Temp workers have therefore no impact on the gap between these two measures. I am not aware of comprehensive firm-level data on temp workers, but according to surveys they typically amount to a few percents, See Havez 2018 for details on those sectors who employ a lot of temporary workers.

[33]For instance, starting from 2004, unemployment benefits received in year n appears as job spells if the individual is employed either in year n+1, or n-1 . This addition was made to allow researcher to track movements in and out of the labor force.

[34]"légère sous-estimation du volume d'emploi total tel qu'il est pris en compte dans la masse salariale" p.154 Beguin and Haag 2017

the Labor Code is the responsibility of the Corps des Inspecteurs du Travail (Workplace Inspectors Unit), a specialized group of civil servant. Its function is similar to the Office of Inspector General in the US Department of Labor. A public report is produced every year at the intention of the International Labor Organization. The impression that emerges is that while these controls are resented by the firm managers, the audits' frequency is low and monetary sanctions for failing to comply with these regulations are uncommon.

First, the average effective number of firms to be controlled by each controller is large. In the2001, they were on average $1,159$ establishments and $11,521$ employees per agent of the follow a rigid division of labor following the seniority of the public servant. Junior employees are called "controllers" and are in charge of establishments with less than 50 employees while the more senior "inspectors" are in charge of those above the threshold. Breaking up the workload according to this line, there were on average $1,667$ establishments and $8,964$ employees per controller and 110 establishments with $16,262$ employees per inspector.

However, the regulations activated at the thresholds are only a tiny fraction of the numerous provisions in the Labor Code. Workplace inspectors must also investigate[35]: infraction to workplace safety and especially workplace injuries, illegal hirings of undocumented workers, the respect of work duration laws, timely payment of wages, compliance with collective wage bargaining, discriminations, sexual and moral harassment, human trafficking, smoking interdiction. Given the large scope of their duties, priorities are defined by the Minister in charge of Labor relations [36] and other potential violations – among which some of the regulations this study focuses on – receive far less attention.[37] For instance, the construction is over-represented. Because a status of limitation of three years applies to many of these infractions, in principle establishments must be controlled at least once every three years. No more than 2% of the cases opened lead to formal indictments or convictions. Most of the times, controlled firms are given a warning and ordered to comply. In theory, penalties for

[35]Article L8112-2 of the Labor Code

[36]Article L8112-1 of the Labor Code

[37]In theory an inspector that notices the unlawful absence of a workplace safety committee is mandated to order the firm to institute one.

infractions to the Safety Code can go up to 3750 euros.[38]

For some of the regulations, direct evidence of compliance can be gathered from surveys run by the administration. I consider two of these surveys: one called REPONSE and one called ACEMO. The first one, REPONSE, is a survey conducted every five years on a sample of about 2% of the firms. The aim of this survey is to better understand the working conditions in the firms and the interactions between workers, workers' representative and management.[39] The survey is addressed to each of these three parties. From this survey, I can test whether some of these policies are followed. If a firm fails to put in place a profit-sharing scheme, employees can ask the judge to impose a default one specified by law.[40]

Finally, empirical studies have found that individuals can overestimate the cost of evasion (Andreoni, Erard, and Feinstein 1998). It is moreover difficult to estimate the cost of the penalties given that an infraction has been detected. Moreover, the great majority of investigations for financial fraud do not give rise to prosecution or end in confidential settlements with the administration.

Evasion is a often serious concern in bunching designs, and third-party sources are held more accurate (see Fack and Landais 2016 or Carrillo, Pomeranz, and Singhal 2017). The true impact of regulatory distortions depends on the extend of the economic response, purged of the evasion behavior. In order to separate the two, I enriched the basic model with an evasion margin, allowing firms can divert resources to engage in misreporting and operate above the threshold without complying with the regulations. This section provides evidence that many firms bunching around the threshold are in fact misreporting and then shows how to leverage the two datasets to infer the perceived cost of the regulation. Since firm owners have a incentive to stay below the threshold for tax purpose, they could be tempted to misreport

[38]Perhaps not surprisingly, the 50-employee threshold also serves for the division the labor among labor inspectors. It is a sensible division since different sets of policies apply below and above that threshold, however it is not particularly helpful to catch misreporting firms. Firms with 52 employees who misreport and declare 48 employees fall under the purview of a more junior controller.

[39]See Appendix for details

[40]see decision Cass. Soc. n° 03-03-10502, 13/09/2005

the number of workers or hire undeclared workers, in which case the sharp discontinuity exaggerates the true allocative distortion. To assess the extend of misreporting, we confront two datasets: a fiscal dataset where firm owners (or, for that purpose, their accountant) report the number of employees on their corporate tax form, and the one generated from payroll data, used to compute payroll taxes as well as entitlements (unemployment and retirement benefits).

Another reasonable concern raised in previous studies that attempted to estimate the impact of the threshold relates to the timing of the labor measures. In principle whether the labor variable in the Social Security data .[41] Payroll data employment figures are the ones coming from other sources and this is true as well in France (Ceci-Renaud and Chevalier 2010). Failure to account for this misreporting channel leads to serious biases. For instance, in an unpublished working paper Smagghue 2014 estimates an elasticity of 0.6 based on the fiscal data. Similarly, Garicano, Lelarge, and Van Reenen 2016's analysis is based on the tax data. What they interpret as investment per capita spikes before the regulatory threshold are actually a spurious consequences of the denominator lowered by the under-reporting behavior of the firms before the threshold.

1.4.8 Missing data.

I also verified that the reporting gap is not due to missing data. About 0.4% of all the firms in the tax dataset lack a firm identifier. For this reason we are unable to match them with their corresponding Social Security data and cannot ascertain whether they are misreporting. This identifiers are missing despite numerous procedures that the administration to undertakes to check the data and retrieve partly missing or erroneous firm identifiers in a separately maintained exhaustive list of firms called "Fichier des Redevables Professionnels de la DGI"(Beguin and Haag 2017). I check that these missing observations have very little impact on our measures of firm size distribution. For instance, in 2004, there is exactly 1 firm with 49 that could not be identified, and 3 firms with 48 employees. Around the 10-employee threshold,

[41]"recalé sur la date de fin d'exercice et transformé, en ce qui concerne l'emploi en nombre de postes, en un concept identique à celui des liasses fiscales"

the absolute numbers are respectively 168, 174, and 103 firms of $8, 9$, or 10 employees. Only at the 20 threshold do we observe some bunching before the threshold: 17 (145), 18 (252), 19 (456), 20 (25).

1.5 Estimation of Firm-level elasticities

This section details the estimation of the sufficient statistics needed to estimate the micro-elasticities and present the baseline results.

1.5.1 Bunching and Counterfactual distribution.

I construct a counterfactual distribution by fitting a 6th order polynomial outside of the bunching region, following the standard approach in the bunching literature (Saez 2010). The magnitude of the bunching – the so-called "missing mass"– is then computed as the deviation from the polynomial fit in an interval around the threshold.

I proceed in a similar fashion and obtain B using by regressing the observed distribution on a polynomial of degree 6, and excluding the region around the threshold. This corresponds to running the following regression

$$h(n) = \sum_{p=1}^{6} a_p \cdot n^p + \sum_{j=N-\delta^-}^{N+\delta^+} d_j \cdot 1_{\{n=j\}} + \xi_n \tag{1.19}$$

where the dummies d_j are just here to absorb the firm counts in the region of the firm size distribution that is distorted by the policy. The counterfactual size distribution is is then estimated by

$$\hat{h}(n) = \sum_{p=1}^{6} \hat{a}_p \cdot n^p$$

Following the methodology that is now standard for the bunching literature (cf. Henrik Jacobsen Kleven 2016), I set the lower bound of the bunching region, δ^- , at the point where the distribution starts increasing. To obtain the upper bound δ^+ I conduct an iterative procedure that ensures that the two distributions – the actual one and the counterfactual one – add up to the same number of firms in the interval considered. The upper bound is such that the displaced mass of firms before the threshold and after the threshold exactly

compensates. Typically because the deviation is much larger before than after the threshold, this method implies an asymmetry in the response region: the upper bound δ^+ is further from the threshold than the lower bound δ^- is. In practice, I cannot equalized the excess mass and the missing mass exactly perfectly because of the round numbers of employees. In practice, I pick the closest value and conduct robustness checks with the other value.

The detailed estimation goes as follow. First, I apply the standard bunching procedure to recover the firm size distribution, the bunching mass $\hat{B}$, the missing mass $\hat{M}$, and the boundaries of the excluded region δ^- and δ^+. As explained above, this step involves iterating until δ^+ is such that the bunching mass and the missing mass coincide, that is, such that the counterfactual firm distribution and the empirical firm distribution integrate to the same number of firms. For the nonparametric estimation of the firm size distribution, I select the best fit among the space of 6^{th} order polynomials as is standard in the bunching literature. This gives me the distribution of productivity up to a scale factor. I then compute the distortions

The standard errors are obtained by boostrapping. More precisely, I re-sample the residuals $\{\xi_i\}$ in equation (1.19) 200 times . The standard errors I then report are the standard deviation of the distribution of estimates. It is important to recognize that because our data consists in the whole universe of firms, – up to at tiny fraction of missing data – these standard errors do not correspond to sampling errors per se but are a inform us about misspecification errors.

The bunching mass is the difference between the empirical distribution and the counterfactual

$$\hat{B} = \sum_{j=N-\delta^-}^{N-1} d_j - \hat{c}_j \tag{1.20}$$

Similarly the missing mass is sum of the difference between the counterfactual number of firms after the threshold and the observerd number of firms.

$$\hat{M} = \sum_{j=N}^{N+\delta^+} \hat{c}_j - d_j \tag{1.21}$$

It is convenient to write these quantities as ratios

$$\hat{b} = \frac{\hat{B}}{\sum_{j=N-\delta^-}^{N-1} \hat{c}_j} \ , \ \hat{m} = \frac{\hat{M}}{\sum_{j=N}^{N+\delta^+} \hat{c}_j} \tag{1.22}$$

The fraction of firms that do not adjust (Henrik J Kleven and Waseem 2013) is $1 - \hat{m}$. In a version of the model with fixed cost of adjustment in capital, the interpration is that a fraction $1-\hat{m}$ of firms are constrained. At the first order, the bunching mass is $\hat{B} \approx s(N) \Delta N$ where $s(N)$ is the empirical firm size distribution. Therefore an estimate for ΔN that can be used in the formulas is $\Delta N \approx \frac{\hat{B}}{s(N)}$. With the reduced-form frictions, $\hat{B} \approx \hat{m}.\, s(N) \Delta N$. Therefore an estimate for ΔN that can be used in the formulas is $\Delta N \approx \frac{\hat{B}}{\hat{m}.s(N)}$. The presence of frictions implies that the underlying structural elasticity could be higher than the one estimated from the response.

1.5.2 Capital distortions.

I now turn to the measures of capital distortions. To control for measurement errors in capital and potential downward biases, I consider several measures of the capital distortions. In theory, one should look at the top buncher, for which the capital distortion reaches its maximum. In practice I implement three estimation strategies, one that uses the average capital distortion in the bunching region, one that exploits the cross-sectional distortion in capital for the bunchers and one that uses investment rates. The model presented in the theoretical framework is extended to accommodate a capital-labor ratio that is increasing with firms' sizes, as it is the case in the data. The firm-level production can be written instead

$$F(K, L) = \left[\alpha(z) K^{\frac{\sigma-1}{\sigma}} + (1 - \alpha(z)) L^{\frac{\sigma-1}{\sigma}}\right]^{\frac{\sigma}{\sigma-1}}$$

Let $\kappa(z) \equiv \frac{\alpha(z)}{1-\alpha(z)}$, then the capital labor ratio outside of the bunching region is $k(z) = \left(\kappa(z) \frac{w}{r}\right)^{\sigma}$. And $\kappa(\cdot)$ can be recovered from the empirical relation between capital intensity and size.

The average capital distortion $\mathbb{E}[\Delta k]$ can be readily measured as the deviation in the capital intensity observed in the bunching region. I test for the presence of capital intensity

distortions by running the following regression at the sector level

$$k_i = \beta_0 + \beta_1 . L_i + \beta_2 . (L_i)^2 + \sum_{j=N-\delta^-}^{N} \delta_j \cdot 1_{\{L_i=j\}} + \epsilon_i \tag{1.23}$$

I use the first two moments of labor to absorb trends in the capital intensity. The dummies interacted with size capture the deviation in the capital intensity. There are several reasons to think that this measure understates the top buncher's distortion. First, each coefficient δ_j in the bunching region is an average response weighted over an interval of firms, as

$$\delta_j \equiv \mathbb{E}\left[\Delta k \,|L = j\right] = \frac{1}{H\left(\bar{z}\right) - H\left(\underline{z}\right)} \int_{\underline{z}}^{\bar{z}} \Delta k\left(z\right) h\left(z\right) dz$$

At the first order approximation, we have if the distribution is taken to be roughly constant over the interval $[\underline{z}, \bar{z}]$, in which case $\mathbb{E}\left[\Delta k\right]$ corresponds to about half of the maximum capital distortion. Such an approximation creates a downward bias since $h\left(\cdot\right)$ is decreasing. To control for the downward bias, I can therefore use the the counterfactual density $\hat{h}$ to better estimate the downward bias created by the bias. Second, measurement errors in capital can create a attenuation bias in the measure of the capital gap. As a robustness check, I estimate how the elasticity would change if the distortions were higher. The results are displayed in Table 1.2 and 1.3. . The first table

The second approach exploits the static first order condition in capital and the relationship it predicts in the cross-section of the firms bunching around the threshold. The intuition is that as firms over-accumulate capital and deviate from the non-distorted capital-labor ratio, their efficiency decreases. The main specification I run to capture the relationship between the distorted output and capital is

$$\log k_i = \gamma_0 + \gamma_1 . \log y_i + \gamma_2 . (\log y_i)^2 + \gamma_3 . (\log y_i)^3 + X'\beta + \epsilon_i$$

The third approach addresses potential measurement issues in capital. I use the information contained in the investment variable, which is known to be more accurately measured than the net value of the firm's capital stock(e.g. Collard-Wexler and De Loecker 2016). In order to do so, I map the parameters of the models into investment distortions using a dynamic version of my baseline model, the details are relegated to the Appendix. The basic

intuition is the following. Suppose a one-period time-to-build such has investing I_t units of good at time t deliver an equivalent of installed capital in the next period. The Bellman equation for the firm's manager takes the following form

$$V_t(K;z) = \max_{I,L}(1-\tau)\left[\Pi(K,L) - I\right] + \beta \cdot \mathbb{E}\left[V_{t+1}(K';z')\right]$$

where the law of motion for capital is $K' = (1-\delta)K + I$ and the law of motion for the productivity index is exogenously specified, $z' = G(z)$. The first order condition in investment is therefore simply

$$I(K,z) = \frac{\beta}{1-\tau}\mathbb{E}_t\left[\frac{\partial V_{t+1}(K';z')}{\partial K}|z\right]$$

As long as the mean growth rate of productivity is not too large, the local deviation around the threshold in investment per capita, $\frac{I}{L}$, reveals how much capital distortion firms are willing to operate on when they get stuck at the threshold. Consistent with the previous approaches, I find little evidence of distortion in investment. Figure plots investment per capita around the threshold for various measure of investment per capita: investment in all assets and investment in PPE capital. The two series are remarkably similar. As a robustness check, I also compare with the series obtained after replacing the numerator with the log changes in PPE capital and the one obtained after replacing the denominator with the self-reported measure of labor instead of FTE labor. The log changes in capital do not display any significant up-tick. In contrast – and as expected if self-reported labor under-estimate the true quantity of labor – the last measure display a slight bump before the 50-employee threshold.

1.5.3 Estimation.

I implement this procedure at each of the regulatory thresholds, i.e. at 10, 20, and 50 employees. This allow me to estimate the elasticity at different points of the distribution. Moreover, the number of firms at the the 10 and 20 thresholds are much bigger which considerably increases the power of our estimation. It turns out that the 20-employee threshold create little distortion. I then assume that the 10 and 50- employee thresholds are well-separated so that we can analyze them separately. I first consider the 50-employee where the impact of the

regulations seems the most salient.

1.5.4 Bunching evidence and Firm-level Elasticities.

Looking at the aggregate size distribution I find a share of non optimizers is 89% in the social security data. This high magnitude is similar to what has been found in the literature. Henrik J Kleven and Waseem 2013 estimate that about 90% of workers do not adjust labor supply due to frictions and L. Liu and Lockwood 2015 find that a similar amount of firms do not adjust turnover in the presence of VAT notches.

I first estimate the elasticity around the 50 threshold and implement the method at the sector level, to allow for potentially large technological differences across sectors for each of the two datasets.At the 50-employee threshold, the elasticities range from 0.04 in Retail to 0.21 in the Wholesale and Trade sector. These

1.5.5 Aggregate elasticities.

The micro estimates obtained in the previous sections characterize the *intensive* margin of capital-labor substitution at the firm-level. I how do these micro-level elasticities translate into an aggregate elasticity of substitution for the economy? First, this framework requires the estimation of several other elasticities which measurement might be uncertain. Second the model abstract from This is why we view this exercise mainly as a way to generate reasonable bounds on the aggregate elasticity. I try different alternatives, setting the markups equal to a standard value or computing the markups from the ratio of sales over costs. The results are presented in table Most of the sectoral markups lie between 0.2 and 0.5. Second, the definition of sectors is likely to matter In our data, we consider all 2-digit industries. And we find as a consequence an elasticity around 0.15. Consequently, these results rule out quite clearly the explanation based on falling capital prices.

1.6 A Structural Model with Frictions in Labor Choice and Misreporting

The baseline model presented in the previous section relates the structural parameters and the policy costs to sufficient statistics. This tractable approach yields closed-form formulas at the expense of abstracting from some potentially important margins. I therefore complement this estimation strategy with a more structural approach. However, as is common with notch designs, the model delivers predicts that there should be a hole in the firm-size distribution right after the threshold, since the decision to locate in this region is strictly dominated. Empirically, the missing mass of firms in the dominated region is between 5 and 10 times smaller than predicted by the simple model. In order to explain the absence of such a "hole", I develop a more structural approach. I then add two margins: optimization frictions and the possibility of misreporting the amount of employees. The optimization friction margin accounts for the mass of firms located after the threshold while misreporting is necessary to account for the gap between the labor measures coming from the two datasets. From revealed preferences, what matters for the input choices is the perceived cost of the policy by the firm manager. The large response in the dataset where the labor measure is subject to manipulation clearly exclude the possibility that firms are unaware of the policies. This framework is built to provide an estimate of the elasticity of substitution that is as reliable as possible and tries to limit misspecification error by placing only limited parametric assumptions. The drawback of the approach is that we have to resort to numerical method and cannot obtain the closed-form expressions or maximum likelihood estimators that one could obtain with stronger parametric assumptions.[42]

Substantial biases in the estimates of the elasticity and of the policy cost are created by failing to take into account that firms might misreport their labor size. Under-reported labor results in an overstated economic response, which creates an upward bias on the elasticity. The investment bump reported in Garicano, Lelarge, and Van Reenen 2016 and in Smagghue

[42]For instance Garicano, Lelarge, and Van Reenen 2016 assume that the unobserved distribution of firms's productivities follows a Pareto distribution but their likelihood estimator is uninformative about the elasticity parameter as it is completely flat on this parameter's dimension (p.8 of their Appendix).

2014 appears as an artifact of firms misreporting their true number of employees. Investment per capita looks higher in the bunching because the denominator is under-reported in this region. The reason is that the presence of firms in the "hole" could be due to frictions that interfere with manager's decisions to bunch. If this is the case, only part of the theoretical, structural response is observed. It is widely recognized that firms face frictions in employment choices (R. Cooper, J. Haltiwanger, and Willis 2007). In the French labor market, often presented as one of the least flexible (Blanchard, Nordhaus, and Phelps 1997) these frictions are more prevalent than in most developped countries. I therefore recognize in the model that managers can only choose imperfectly their desired employment level and face optimization frictions. I introduce noise in the optimization program of the manager as in Chetty 2009.[43] This is just one parsimonious way of capturing the effect of several labor frictions that have been examined in the literature such as search frictions or economic uncertainty. If L is the targeted employment level, then the realized employment is given by $\tilde{L} = L \cdot \exp(\epsilon)$, where $\epsilon \sim N(0, \theta)$. As a consequence, the annual Full Time Employment count differs from the equilibrium value chosen *ex ante*.

The timing of the manager's decision is the following: at the beginning of the period, the manager chooses the optimal amount of capital, then optimal chooses a target labor, then the uncertainty is realized. Given a choice of capital, the firm's objective function becomes

$$\begin{aligned}
\Pi_K(L; z, K) &= \mathbb{E}_\epsilon \left[\left(1 - \tau - \Delta\tau . 1_{\{L\exp\{\epsilon\} \geq N\}}\right) \left(z^{1-\nu} F(K, L\exp\{\epsilon\})^\nu - wL\exp\{\epsilon\} - rK - c(s)\right) \right] \\
&= (1-\tau)\, \mathbb{E}_\epsilon \left[z^{1-\nu} F(K, L\exp\{\epsilon\})^\nu - wL\exp\{\epsilon\} - rK \right] \\
&\quad - C(L; \theta, \tau) \\
&= (1-\tau)\, z^{1-\nu} \mathbb{E}_\epsilon \left[F(K, L\exp\{\epsilon\})^\nu \right] \\
&\quad - wL\exp\{\theta^2/2\} - rK - C(L; \theta, \tau)
\end{aligned} \tag{1.24}$$

where $C(L; \theta, \tau)$ is the expected cost for the firm of crossing the threshold and having to

[43] An alternative hypothesis would be to suppose that managers can be of several types: "good" or "honest" managers who would never cheat and "dishonest" managers who are prone to cheating and misreporting their size.

implement the regulations. It is defined as

$$\begin{aligned} C(L;\theta,\tau) &= \Delta\tau.\mathbb{E}_{\epsilon}\left[z^{1-\nu}F(K,L\exp\{\epsilon\})^{\nu} - wL\exp\{\epsilon\} - rK \mid L\exp\{\epsilon\} \geq N\right] \\ &\qquad \times \Pr\{L\exp\{\epsilon\} \geq N\} \\ &= \Delta\tau.z^{1-\nu}\int_{\ln\frac{N}{L}}^{\infty} F(K,L\exp\{\epsilon\})^{\nu}\phi(\epsilon)\,d\epsilon \\ &\qquad -\Delta\tau w - \Delta\tau.rK\left[1-\Phi\left(\ln\tfrac{N}{L}\right)\right] \end{aligned} \tag{1.25}$$

where $\Phi(.)$ is the cdf of a the standard normal distribution. The labor friction has several implications on the profit function and on the firm's behavior in equilibrium. The main distinction with the baseline model is that the profit function is no longer discontinuous, as the labor friction smooths out the discrete cost of crossing the threshold. Moreover, the uncertainty costs increase with the amount of noise.

Proposition 4. *The expected cost of the regulation increases with the amount of uncertainty, i.e.* $\frac{\partial C}{\partial \theta} > 0$.

The intuition is that as the labor friction because larger, firms'managers have an incentive to target an employment level further below the threshold to reduce the likelihood of paying the regulation cost.

1.6.1 Misreporting.

In the data, many firms located around the threshold seem to be misreport their true amount of labor. I therefore add an evasion margin in the model. As in Chetty 2009, I suppose that hiding a fraction s of their labor force is costly and that firms must divert an amount $c(s)$ of resources . One interpretation of that cost is that to bypass the regulation manager need to divert time or pay an accountant to help with the dissimulation. Finding ways to evade the regulation might create all sorts of inefficiencies. Managers must also balance the benefits of evading with the expected cost of being caught. In that case one can write $c(s) = p(s)\cdot t(s)$, where $p(s)$ is the perceived probability of being caught and $t(s)$ the associated fine. I follow Chetty, Friedman, et al. 2011 in assuming that the evasion cost $c(.)$ is nondecreasing and strictly convex. A parsimonious parametrization is to consider the special case: $c(s) = \frac{s^{\gamma}}{\gamma}$,

following Chen et al. 2017. More general shapes of misreporting cost could be considered. Ultimately what matters for the purpose of this article is that the convexity ensures that misreporting arises as local phenomenon around the threshold. The realized employment is again given by $\tilde{L} = L \cdot \exp\{\epsilon\}$, where $\epsilon \sim N(0, \theta)$ however the reported labor is now different from the actual labor and depends on the amount of misreporting $\tilde{L}_{rep} = L \cdot \exp\{\epsilon - c(s)\}$. The cost of the regulation still applies to the actual workforce but its cost is borne only if the reported labor force exceeds the threshold. [44]The timing of the decisions is the same as above, after having chosen K, the profit is

$$\begin{aligned}
\Pi_K(L; z, K) &= \mathbb{E}_\epsilon\left[\left(1 - \tau - \Delta\tau.1_{\{Le^\epsilon \geq N\}}\right)\left(z^{1-\nu}F(K, Le^\epsilon)^\nu - wLe^\epsilon - rK - c(s)\right)\right] \\
&= (1-\tau)\, z^{1-\nu}\mathbb{E}_\epsilon\left[F(K, L\exp(\epsilon))^\nu\right] - c(s) \\
&\quad -wL\exp\left(\theta^2/2\right) - rK - C(L; \theta, \tau)
\end{aligned} \tag{1.26}$$

where the expected cost of the regulation is

$$\begin{aligned}
C(L, s; \theta, \tau) &= \Delta\tau.\mathbb{E}_\epsilon\left[z^{1-\nu}F(K, Le^\epsilon)^\nu - wLe^\epsilon - rK - c(s) \mid Le^{\epsilon - s} \geq N\right] \\
&\quad \times\left[1 - \Phi\left(\ln\tfrac{N}{L} + s\right)\right]
\end{aligned} \tag{1.27}$$

It corresponds to the expected fraction of the profit that will be dissipated by the hassle cost when the labor count crosses the threshold. In equilibrium, the incentives to misreport balance the expected reduction in the burden created by the regulation and the first order condition in the amount of misreporting is

$$c'(s) = -\frac{1}{(1-\tau)}\frac{\partial C(s, L; \theta, \tau, \phi)}{\partial s}$$

Misreporting workers, reduces the expected burden because it correspond to a downward shift in the distribution of the labor shock. Alternatively, the effect of misreporting is

[44]There could be at least two sources for these costs: a penalty after a control by the administration and deteriorated working relations with the employees. Section 3 provides details regarding the enforcement. It is also possible that managers overestimate the costs of non-compliance (Andreoni, Erard, and Feinstein 1998) either by overestimating the detection probabilities or the penalties associated with tax evasion (see Scholz and Pinney 1995; Chetty 2009).

equivalent to raising the threshold above which the regulations must be enforced, which reduces the probability of being subjected to the regulation. Firms that report a workforce above the threshold do not misreport neither do those that are far below the threshold. The possibility of misreporting creates an additional distortionary effect on the marginal product of labor. This distortion is negligible away from the threshold as the uncertainty of whether the threshold is crossed the threshold or not vanishes. However in the region before the threshold, the uncertainty has a "shading effect" on the choice of labor as firms cut down on labor to limit the risk of going above the threshold. In this structural model, the amount of bunching firms is influenced by three effects: i) the substitution between capital and labor ii) the misreporting of the true amount amount of employee iii) the "shading" effect to prevent against crossing the threshold because of the noise. Because the model is not linear, these effects interacts and it is not possible to separate them in closed-form expressions. The misreporting channel exaggerates the amount of bunching in the self-reported data and the shading turn the sharp bunching more into a smooth bump, which better corresponds to the pattern in the matched employer-employee data.

I estimate the structural model using a simulated method of moments. I simulate the model and require the simulated moments to match the size and capital distortion. More formally, my estimator is chosen to minimize the vector of moments

$$\hat{\beta} \in \arg\min_{\beta \in B} \left(\mathbf{m} - m\left(\beta\right)\right)' W \left(\mathbf{m} - m\left(\beta\right)\right)$$

where m is the vector of moment and W is a weighting matrix that is set initially to be the identity matrix. I then compute the optimal weighting matrix W_n by taking the inverse of the variances of the empirical moments at the minimum. Like in Einav, Finkelstein, and Schrimpf 2015 I include in those moments the histogram of the distribution around the threshold , the excess mass, and the gap between the two measures. In effect, I use the bunching equations as an auxiliary model to provide further moments and finds the parameters that minimize the moments. For robustness checks, I use alternative measures of capital: total net assets, net tangible capital and gross tangible capital. Using the two datasets for the estimation increases the power of our estimates as it provides cross-equations

restrictions. For instance the profile of the gap between the reported measure depends on both the regulation cost and the misreporting cost. The moments need to be carefully chosen in order to minimize misspecification errors while capturing the essential economic features. Some of the earlier works mentioned previously lead to very wide standard errors for the elasticity estimate. I do not make parametric restrictions on the underlying, unobservable distribution of productivities and instead choose a flexible polynomial approach.

Besides the elasticity of substitution between capital and labor, σ, we have four other parameters that we need to estimate: the variance of the optimization error (θ), the convexity of the misreporting cost function (γ), the returns to scale (ν), and the capital share coefficient in the CES (α). The convexity of $c(.)$ is obtained from matching the average gap between the reported employment and the social security measure. More precisely, we reproduce the increase in the gap around the threshold. The magnitude of the returns to scale calibrated to range commonly found in the literature. I pick a baseline value of $\nu = 0.85$ as in Atkeson and Patrick J. Kehoe 2005 and conduct sensitivity analyses around this value, in the range suggested by Basu and Fernald 1997. An alternative estimation approach would be to use directly the first order conditions and derive their implications on factor ratios. Such an approach raises at least three concerns: i) it first implies to make assumptions about r , the rental rate of capital, for instance following Hsieh and Klenow 2009, and ii) measurement errors in capital can create severe bias (see Collard-Wexler and De Loecker 2016). For the baseline tax rate τ, I use the prevailing statutory corporate income tax rate in each year. During the period considered, the smallest firms can qualify for a "reduced tax rate". To qualify for this reduced rate, the firm's revenues must be below a certain threshold (about 1 million euros[45]). The reduced rate then applies to the first bracket of the revenues. For all the revenues above this bracket, the normal rate applies. Because this bracket is very low – less than $50k$ euros – it is not directly relevant for this quantitative exercise and I do not model it explicitly.

The results of this structural estimation are in line with the results based on the reduced-

[45]See General Tax Code (CGI) article 219, I-b

form formulas. In the structural estimation, I consider only the 50-employee threshold and I assume, as in Henrik J Kleven and Waseem 2013 that the other thresholds are far enough that they do impact the behavior of the firms around the 50-employee threshold. In Table 4 I present the results I obtain for the manufacturing sector. The cost of regulation that is about 0.12% of the profits and the elasticity is 0.18, somewhat higher than what estimated with the first approach.

1.6.2 Discussion and implications.

How do these estimates help us understand the evolution of factor shares? Between 1980 and 2010, the French labor share fell by 7.3 percent points, a fall of similar magnitude than the decline of the U.S. labor share's over the same period (7.5 percentage points). However, a closer look at their evolutions over the three decades reveal strikingly different very patterns. The U.S. labor share was flat roughly until 2000 and then dropped sharply (Figure (1.14)).[46] In France, the pattern in almost the complete opposite: a brutal drop from 1983 to 1989 and then a flat if somewhat upward trajectory. A low elasticity leaves little scope for factor prices to affect the labor share. Even if one takes the view that my estimates are mostly reflecting the short-run response. It is highly unlikely that such a sharp drop can be accounted for by entry and exit patterns. [47]

A low elasticity has several other important implications. For instance, in a static input-output model as in Baqaee and Farhi 2017 one can show the impact of a productivity shock in sector $k$on aggregate consumption depend on a combination of firm-level elasticities and markups . With an elasticity smaller than one, the shocks to productivity are *attenuated*.

[46]The U.S. labor might have fallen by 2 percentage points more if one include the fact that U.S. multinational firms have increasingly shifted part of their domestic profits abroad. See Guvenen et al. 2017 for the impact of offshore profit shifting on U.S. national accounting.

[47]I measure the labor share as the "corporate labor share", using the definition and data provided by Karabarbounis and Neiman 2014. The corporate labor share excludes government activities as well as the labor and capital income earned by unincorporated businesses or sole proprietors, for which the labor share needs to be imputed. Thus the corporate labor share is a metric that is less sensible to these imputations methods and more robust for international comparisons. The French corporate labor share fell from 73.4 % to 66.1% while the U.S. labor share fell from 64.7% to 57.2%.

With elasticities close to 0, these shocks barely propagate. Therefore low elasticity limit some of the propagation channels of production-network macroeconomic models.

How would adding adjustment costs to capital affect my elasticity estimates? While labor frictions are present in my structural framework and help account for the sizable amount of firms lying in the dominated region, inputs in my models are static and I abstracted from the full dynamics of firms' input decisions. However, because some firms in the bunching region have been here for several years, my estimate captures also this medium-run response. For instance, about 20% of the firms that were bunching at 50 in 2000 where still bunching there 7 years later (see Figure 1.11). Quadratic adjustment costs in capital of the type popularized by Sargent 1978 create second order disturbances for the relatively small investment deviations considered here, and are therefore unlikely to generate any substantial downward bias. More critical are fixed cost of adjustment, which generate the large inaction bands characteristic of S-s models (Ricardo J Caballero and E. M. Engel 1999). An important body of literature argue that fixed costs of adjustment are necessary to match the moments of the investment distribution at the plant-level (R. W. Cooper and J. C. Haltiwanger 2006). However the magnitude of these fixed costs is hard to pin down precisely and seems sensitive to modeling choices. For instance, Khan and Thomas 2008 and Bachmann, Ricardo J. Caballero, and E. M. R. A. Engel 2013's estimates are one order of magnitude apart and, moreover, Khan and Thomas 2008 show that the plant-level investment cross-section is consistent in general equilibrium with negligible fixed costs. If their view is correct, then one can reasonably expect my elasticity estimates to be largely unaffected by adding these fixed costs. Otherwise, my elasticity estimates need to be interpreted as the *effective* elasticities of substitution, corresponding to the firms' reactions in the face of lumpy adjustment.

Could a more explicitly dynamic model of the bunching behavior deliver substantially higher elasticity estimates? There are good reasons why this would not necessarily the case, as several forces play against a high elasticity. If anything, my static design suffers from an overestimation bias. Because static designs pool together cross-sections from different years, firms who bunch in consecutive years will be counted several times in the excess mass. Those "repeat" bunchers who stay at the threshold several years in a row tend therefore to

inflate the excess mass in the bunching region. By overweighting those repeat bunchers, the estimate produced will overestimate the true elasticity. To exploit the panel dimension in my setting I can replace my identification assumption that the counterfactual employment distribution is smooth by the assumption that the counterfactual employment *growth rate* distribution is smooth. By correctly re-weighting the contribution of repeat bunchers, Marx 2015 shows that the overestimation bias can be up to an order of magnitude. Because my baseline estimate for the elasticity is already low, this potential upward bias does not present a fundamental challenge to my conclusions.

Finally, this paper has so far concentrated on the *intensive* margin of substitution between capital and labor. Indeed, the static framework I develop abstracts from entry and exit of firms and from technological change. An active extensive margin would provide a mechanism through with the long-run elasticity could diverge from the short-run elasticity. For instance, if technological change is embodied into capital, then the entry of firms with capital intensity consistently higher than the incumbent's capital intensity would progressively raise the long-run aggregate capital labor ratio. This mechanism is best analyzed in a model with putty-clay technology.[48] When the production function is putty-clay, capital is malleable *ex ante* but once the capital intensity is chosen, it must stay fixed forever. For instance the production function can be CES *ex ante*, but becomes Leontief *ex post*. The speed at which the short-term elasticity converges to the long-run elasticity hinges on the amount of churn in the economy.[49] With entry rate in France slightly below than 9%, and concentrated among the smallest firms (less than 10 employees) this channel is quantitatively limited in my setting. Moreover, since my estimation pools firms across years – and some of these firms have been staying at the threshold for several years – this adjustment effect is already partly reflected in my estimates.

[48]The seminal contribution by Johansen 1959 launched a large literature investigating the property of growth models with putty-clay technology (Calvo 1976; Cass and Stiglitz 1969; Phelps 1963; Solow 1962). More recently, putty-clay model have been used to investigate the evolution of factor shares (Ricardo J Caballero and Hammour 1998), energy use (Atkeson and Patrick J Kehoe 1999), business cycle dynamics (Gilchrist and Williams 2000), or stock market volatility (Francois Gourio 2011).

[49]Sorkin 2015 develops a model along these lines in order to study how employment respond over time to changes in the minimum wage.

1.7 Concluding Remarks

This paper argues that the aggregate elasticity of substitution between capital and labor might be lower than conventionally assumed. The main reason is that the firm-level elasticities are close to zero, putting it differently, the production structure of firms is sticky or rigid, similar to a Leontief technology. The finding is robust across sectors. I develop a new method to estimate this micro elasticity using the distortions generated by size-dependent policies in France. I implement this method on two comprehensive administrative datasets. Despite the striking discontinuity visible in the French firm size distribution, most of the bunching response is in fact due to misreporting by firms. I identified this misreporting behavior using the gaps between the self-reported measure in one dataset with a measure verified by the administration using employer-employee data. These regulations are salient enough that they elicit a significant response at the threshold – even when misreporting is taken into account – , but they generate little distortions in the use of production factors. In particular, distortions in the capital intensity at the threshold or in investment behavior are about 4 to 5 times smaller than what would be implied by a Cobb-Douglas production function. I then aggregate these firm-level elasticities in a framework that allows for substitution across firms and sectors.

More general lessons can be gleaned from this exercise. A low elasticity of substitution between capital and labor has important consequences for investment policies and tax incidence. My findings imply that is that little aggregate capital-labor substitution is the result of the intensive margin. If one takes the view that the aggregate elasticity is large, then the a stronger emphasis should be put on the extensive margin, that is, the entry of new firms and the exit of failing firms. If new firms embody technological changes, then a dynamic entry margin can alter the capital labor mix in the medium run. Alternative explanations might account for the drastic fall in the labor share observed in some countries, such as drastic institutional changes or macroeconomic shocks. My empirical design, which by its very nature focuses on small and local changes, is silent about the role of these institutional changes. Examining in greater depth and through the lens of firm-level data how new entrants con-

tribute empirically to the aggregate elasticity of the economy is an interesting question for left for future research.

1.A Mathematical Appendix

1.A.1 Alternative models of the cost of regulation

In order to capture the large set of regulations that kick in at the regulatory thresholds, the main model analyzes their impact if they operate as a wedge on the tax rate. In order to asses the robustness of this assumption, this section analyzes alternative specifications of the policies – as a "pure" notch, as a combination of a labor wedge and a fixed cost, and as a wedge on output. I derive the bunching formulas in each case. These formulas follow from the profit indifference condition and capture the trade-off between the regulation cost or the cost of distortions around the threshold.

"Pure" Notch

Instead of modelling the cost of the regulations as an increment in the tax rate, $\Delta\tau$, one can view it instead as a fixed cost ΔT that creates a "pure" notch following the terminology of Henrik J Kleven and Waseem 2013. The profit function for the baseline model then becomes

$$\Pi(L;z) = \begin{cases} (1-\tau)\left[z^{1-\nu}L^{\nu} - wL\right] & \text{if } L < N \\ (1-\tau)\left[z^{1-\nu}L^{\nu} - wL\right] - \Delta T & \text{if } L \geq N \end{cases}$$

For this case with no substitution, the bunching formula is

$$\Delta T = (1-\tau)\left(\frac{\nu}{w}\right)^{\frac{\nu}{1-\nu}} z\left(1+\frac{\Delta z}{z}\right)\left[1-\left(\frac{1}{1+{}^{\Delta z}/_{z}}\right)^{\nu}\right] \tag{1.28}$$

Labor Wedge and a Fixed Cost

Suppose that instead of incurring a hassle cost, the firm's faces a wedge on the marginal labor cost τ_ℓ and a fixed cost ϕ, as in Garicano, Lelarge, and Van Reenen 2016. The profit function for the case with no substitution, which is the one they consider, is

$$\Pi(L;z) = \begin{cases} z^{1-\nu}L^{\nu} - wL & \text{if } L < N \\ z^{1-\nu}L^{\nu} - w(1+\tau_\ell)L - \phi & \text{if } L \geq N \end{cases}$$

The profit indifference condition delivers the following bunching formula

$$\left(1+\frac{\Delta N}{N}\right)^{1-\nu}-(1-\nu)\left(1+\frac{\Delta N}{N}\right)=\nu\cdot\frac{1}{1+\tau_\ell}\left(1-\frac{\phi}{wN}\right) \tag{1.29}$$

First, it is worth stressing that this formula is independent from assumptions on the underlying productivity distribution. Second, either an increase in τ_ℓ or in ϕ increases the size of the gap, as the left-hand side is strictly decreasing in $\bar{N}$, declining from ν to 0, when the ratio $\frac{\bar{N}}{N}$ increases from 1 to $\left(\frac{1}{1-\nu}\right)^{\frac{1}{\nu}}$ (≈ 9.3 if $\nu=0.85$).

Proposition 5. *The labor wedge and the fixed cost are not separately identified without further assumptions on the unobserved productivity distribution.*

This can be seen directly in (1.29). As long as the pair of costs (τ_ℓ,ϕ) is such that the quantity $\frac{1}{1+\tau_\ell}\left(1-\frac{\phi}{wN}\right)$ stays constant, then the size of the hole stays the same. Indeed this predicted linear relationship holds almost perfectly for the estimates in Table 1 in Garicano, Lelarge, and Van Reenen 2016. Any other moments of the size distribution – the amount of bunching or shifts in the slope of the size distribution – are directly dependent on assumptions on the unobserved productivity. For instance, the amount of bunching depends on the distribution between N and $\bar{N}$ and any shift in the distribution above $\bar{N}$ can be produced either by increasing τ_ℓ or by a rescaling of $h(.)$, which is not observed. Indeed the firm's labor demand

$$L(z)=\begin{cases} z\left(\frac{\nu}{w}\right)^{\frac{1}{1-\nu}} & \text{if } z\leq \underline{z_0} \\ N & z\in\left[\underline{z_0},\bar{z_0}\right] \\ z\left(\frac{\nu}{(1+\tau)w}\right)^{\frac{1}{1-\nu}} & \text{if } z\geq \bar{z_0} \end{cases}$$

implies the following size distribution, in terms of the underlying productivity distribution $H(.)$,

$$s(n)=\begin{cases} \frac{1}{L_0}h\left(\frac{n}{L_0}\right) & \text{if } n\leq N-2 \\ H(\bar{z})-H(\underline{z}) & n=N-1 \\ 0 & N\leq n<\bar{N} \\ (1+\tau)^{\frac{1}{1-\nu}}h\left(\frac{n}{L_0}(1+\tau)^{\frac{1}{1-\nu}}\right) & \bar{N}\leq n \end{cases}$$

where $L_0 = \left(\frac{\nu}{w}\right)^{\frac{1}{1-\nu}}$. With a wedge on labor, the capital-labor ratios are different on both sides of the thresholds, and the relative gap is $\sigma\Delta\tau$. This provides an additional identification equation.

Output Wedge

Suppose that instead of incurring a hassle cost, the regulations act as an output wedge τ_y. One interpretation is that the manager must now divert some of his time to handle the implementation of these regulations. For the sake of notation, I omit the baseline tax rate and the profit function is

$$\Pi(L;z) = \begin{cases} z^{1-\nu}L^{\nu} - wL & \text{if } L < N \\ \left([1-\tau_y]\,z\right)^{1-\nu} L^{\nu} - wL & \text{if } L \geq N \end{cases}$$

The profit indifference condition delivers the following bunching formula

$$\left(1+\frac{\Delta N}{N}\right)^{1-\nu} - (1-\nu)(1-\tau_y)\left(1+\frac{\Delta N}{N}\right) = \nu \tag{1.30}$$

1.A.2 Investment and Capital distortions in a Dynamic Model

Investment is measured with greater accuracy than capital, which is know to be measured with substantial margin of error [50]. Using the distortions in per capita investment for firms around the thresholds to detect distortionary impacts of the policy is a good way of checking the robustness of the results that rely on capital measures.[51] I therefore extend my baseline model to a dynamic setting in order to make explicit the connection between investment distortions and capital distortions as a reaction to the policies. Investment takes one period to be fully operational and the Bellman equation for the firm's manager takes the following generic form

$$V_t(K;z) = \max_{I,L}(1-\tau)\left[\Pi(K,L) - I\right] + \beta\cdot\mathbb{E}\left[V_{t+1}(K';z')\right]$$

[50] See Becker et al. 2006 for survey of the methods to measure capital.

[51] Collard-Wexler and De Loecker 2016 develop a control function approach that uses investment lags as an instruments for capital.

where the law of motion for capital is $K' = (1-\delta) K + I$ and the law of motion for the productivity index is exogenously specified, $z' = G(z)$. The first order condition in investment is therfore

$$I(K, z) = \frac{\beta}{1-\tau} \mathbb{E}\left[\frac{\partial V_{t+1}(K'; z')}{\partial K} | z \right]$$

relationship between investment per capita, $\frac{I}{L}$ and firm size L depends on the specifications of the law of motion of the productivity index z. To build intuition, lets' first consider some simple cases and assume $\beta = 1$. If z grows deterministically at rate g then investment per capita is equal to $\frac{I}{L} = \delta k^* + k^* \cdot g$ for all the firms except for the ones bunching before the threshold. For the firms who expect to bunch in the following period it is optimal to choose a higher investment rate $\frac{I}{L} = \delta k^* + k^* \cdot g + \Delta k$ in order to operate at a higher capital intensity when they are bunching. In this case the distortion in the capital labor ratio is exactly informative of the capital distortion Δk, as in the case of sharp bunching.

Starting from 2000, there are two direct measures of investment in FICUS, one that sum tangible and intangible investments (INVAVAP) and one that measure tangible investments only (INVCORP). In order to better understand investment patterns, I also look at a more granular version of the datasets only available after the 2009 great recession. I consider three broad investment categories: investment in tangible assets (property, plant, equipment, and land), investment in intangible(intellectual property, R&D, licenses), and financial (mostly security deposits and guarantees).

Capital-Skill complementarity The firm-level production functions can be extended in order to allow for a more than two factors. For instance one can think of several categories of labor as in Krusell et al. 2000. The French matched-employee data seems specially appealing since in the dataset, each job spells receives an occupation code classifying this job in a 500+ occupation category. This occupations are used in Caliendo, Monte, and Rossi-Hansberg 2015 to analyze the hierarchical organization of French manufacturing firms. I compute at the firm-level the share of each category. After removing the unemployment and retirement spells, jobs are organized into five broad categories, each of them being itself with subdivisions: CEO, firm owners or firm directors; senior staff or top managers; employees

at the supervisor level which includes quality control technicians, technical, accounting, and sales supervisor; qualified and unqualified clerical employees (secretaries, human resources or accounting employees, telephone operators, and sales employees), and blue-collar qualified and unqualified workers (e.g. welders, assemblers, machine operators, maintenance workers). I group the first three together to form a high-skill category and the last two in a low-skill category. The share of each category in the wages and hours worked is depicted in Figure . There is suggestive evidence that the composition of labor changes slightly around the threshold. Unfortunately, there are many missing occupation labels and, moreover, an abrupt change in the percentage of missing values happens concurrently around the 50-employee threshold. One interpretation is that the extra administrative requirements generated by the regulatory threshold also forces some employer to be more diligent. To measure carefully the changes in labor composition, one would therefore need a deeper investigation to precisely understand how this two effects interact. In a chapter of I leave this for future research.

1.B Data and Institutional Background

1.B.1 Size-dependent Regulations

This section provides details about the list of the regulations activated when a firm crosses the 10-, 20- and 50-employee thresholds. The specifics of these policies have been modified several times and calls for heir suppression or smoothing out have been frequent in policy and business circles. After some changes in 2008 (Loi de modernisation de l'économie du 4 août 2008), and in 2015, some of the threshold-based regulations have been further relaxed in a bill voted in the French National Assembly this Fall.). I focus on the regulations that were in place during the period 2000-2008. These policy thresholds are commonly referred to, in the public debates, as "Social and Fiscal Thresholds" ("Seuils sociaux et fiscaux"). The regulations are scattered across several pieces of legislation: the Labor Code (Code du Travail, CT), the Social Security Code (Code de Securite Sociale, CSS), the Territorial Authorities Code (Code General des Collectivites Territoriales, CGCT), the General Tax Code (Code General des Impôts, CGI). These regulations have been altered over the last

decades, but not in the period I consider. Calls to eliminate these thresholds go as far back as 1984 (Gattaz 1979). A last change happened in 2015 when it was decided to smooth over time the effect of these thresholds. Firms crossing the threshold are given 2 years to comply.

Regulations starting at the 50-employee threshold A complete list of the regulations can be found in Ceci-Renaud and Chevalier 2010. The most important policies that are activated at the 50-employee threshold are: i) an elected workplace committee must meet once every other month. Elections must be held (unless there is no candidate; ii) information must be transmitted quarterly and yearly about the firm's financial situation to the workplace committee; iii) external "representative" trade unions can designated a union representative; iv) a workplace safety committee must be elected, with up to 3 members, which 2 paid hours every month to each member in order to run the committee; v) compulsory annual negotiation about wages, work duration, etc; vi) if the firm increases dividends with respect to the previous two years then a bonus must also be paid to employees; vii) a specific procedure ("Plan de Sauvegarde de l'Emploi") must be put in place a mass layoff is considered (where mass layoff is defined as the layoff of at least 10 employees over less than 30 days); viii) workplace negotiations must be held to discuss about prevention of accident injury and work penibility; ix) a compulsory monthly report of job turnover must be sent to the administration.

Regulations starting at the 10-employee threshold Very small businesses are exempted from many administrative regulations and many exemptions stop when the firm reaches 10 employees. When they cross that threshold, these firms are subject to: i) a professional training contribution amounting to at least 1.6% of the wage bill (Labor Code - Article L6331-9) ii) a transportation contribution (Code général des collectivités territoriales - Article L2333-64); iii) extra administrative requirements: transmission of national unemployment agency of proof of employment termination (Labor Code - Article R1234-9) and rules regarding dismissal are strengthened.

Regulations starting at the 20-employee threshold. At the 20-employee threshold, the main regulations are the following: i) an increase in the professional training contribution from 1.05% to 1.60% of the wage bill, ii) the obligation that employees with disabilities represent at least 6% of the workforce,[52] iii) two extra contributions each amounting to 0.45% of the wage bill; iv) extra cost for overtime work;

1.B.2 Dataset coverage and the universe of French firms.

They are several available administrative firm datasets generated from tax data, following evolutions in their collection and in their statistical treatment. In some years, these datasets overlap. The datasets whose access is given to researchers is not raw. The administration has developed procedures to remove some inconsistencies and errors. Efforts are made to cross-check some of the information (for instance firm identifiers). These procedures have evolved over time. The last version of the cleaning procedure is explained in painstaking details in Beguin and Haag 2017. The dataset initially covered only firms subject to corporate income tax. In FICUS, the coverage has been extended to firms operating under different, usually simplified, tax regimes, such as pass-through businesses, partnerships or medical practices. Because the tax forms differ, some variables for firms reporting under the other tax regimes can only be estimated. In particular, the only balance sheet items reported by firms under the "Non commercial profits" regime (BNC for "Benefices non-commerciaux", about $500,000$ legal entities report under this tax regime in 2005) are gross capital and operating income.[53] To make my work more comparable to previous studies I conduct my main analysis using FICUS, that is the whole universe of French firms combining all tax regimes despite potential heterogeneity concerns. As a robustness check I performed the analysis excluding these other firms.

[52]Typically, a firm with just 20 employees which did not have any employee with disabilities, has to hire one when crossing the threshold.

[53]see Beguin and Haag 2017 p.126

1.B.3 Evidence for Evasion in France

The gap between reported and measured employment – and its implication in terms of misreporting and regulation avoidance – might at first be surprising. However my empirical finding are in line with Ceci-Renaud and Chevalier 2010 and confirmed by Havez 2018.Fack and Landais 2016 finds other evidence regarding lack of enforcement regarding charitable contributions in France, while Spire (2017) marshals detailed evidence that tax evasion is not strongly prosecuted in France. Overall tax evasion rates in France is found to be in line with the tax evasion rates in the US, estimated at around 16% for individual income (Internal Revenue Service 2008) .

1.B.4 Corporate Income Taxation

The French corporate income tax rate has been remarkably stable over the period I study. From 1993 to 2016 the standard tax rate stood at 33,33%. There is also a reduced tax rate that applies for profits up to $38,120$ euros. All profits exceeding that threshold are taxed at the standard rate. To qualify for the reduced rate, the firm's revenues must not exceed $7,63$ million euros. Because such a threshold could be easily taken advantage by large firms splitting themselves into several several units, an extra requirement is that more than 75% of the paid-in capital of the firm must be detained by natural persons. The 7,63 million euros threshold is therefore a kink point. Consistent with many studies which find little bunching at kink points, I find that there is no significant bunching at the 7,63 million euros threshold. As a result the distribution of the average effective tax rate paid by firms has a peak at $33^1/_3$% and a peak at 0% for firms who are not turning profits. There is a small mass of firms lying between 15%, and $33^1/_3$%, and no significant bunching at 15%, the reduced-tax rate.

1.B.5 Boundary of the Firm

I adopt the same unit of observation as the one used in the FICUS dataset, that is, a firm is defined as a legal entity, associated with a unique SIREN identifier. This by far the

most frequent choice, although there are at least three possible levels of observations, at the establishment level, at the firm level, and at the group level. For the manufacturing sector, the establishment level would correspond to the plant level, which is also often used. In the administrative dataset based on the tax returns, firms are uniquely identified by their 9-digit SIREN number. But firms can locate their activities across several establishments. Establishments are uniquely identified by a SIRET number, which corresponds to the SIREN prefixed by an block of 5 digits. In principle, the social security forms are filled at the establishment level[54]. In addition to that, European Law[55] provided in 2008 a definition of an "enterprise", as "the smallest combination of legal units that is an organizational unit producing goods or services, which benefits from a certain degree of autonomy in decision-making, especially for the allocation of its current resources. An enterprise carries out one or more activities at one or more locations. An enterprise may be a sole legal unit." This definition has been transposed into French Law in 2008 by the "Loi de Modernisation Economique". The French statistical agency is in the process of constructing comprehensive datasets that reflects this new definition. This is a difficult process that requires additional and detailed information about intra-group transactions. Only 120 such "enterprises" have been constructed in the 2015 datasets out of about 80,000 enterprises. Therefore I leave for future research a detailed exploration of capital-labor substitution within the boundaries of these enterprises.

[54]see Beguin and Haag 2017.

[55]Council Regulation 696/93, https://publications.europa.eu/en/publication-detail/-/publication/1ea18a1a-95c2-4922-935c-116d8694cc40/language-en

1.C Additional Figures

Figure 1.3: Effects of the Policies on Labor, Capital intensity

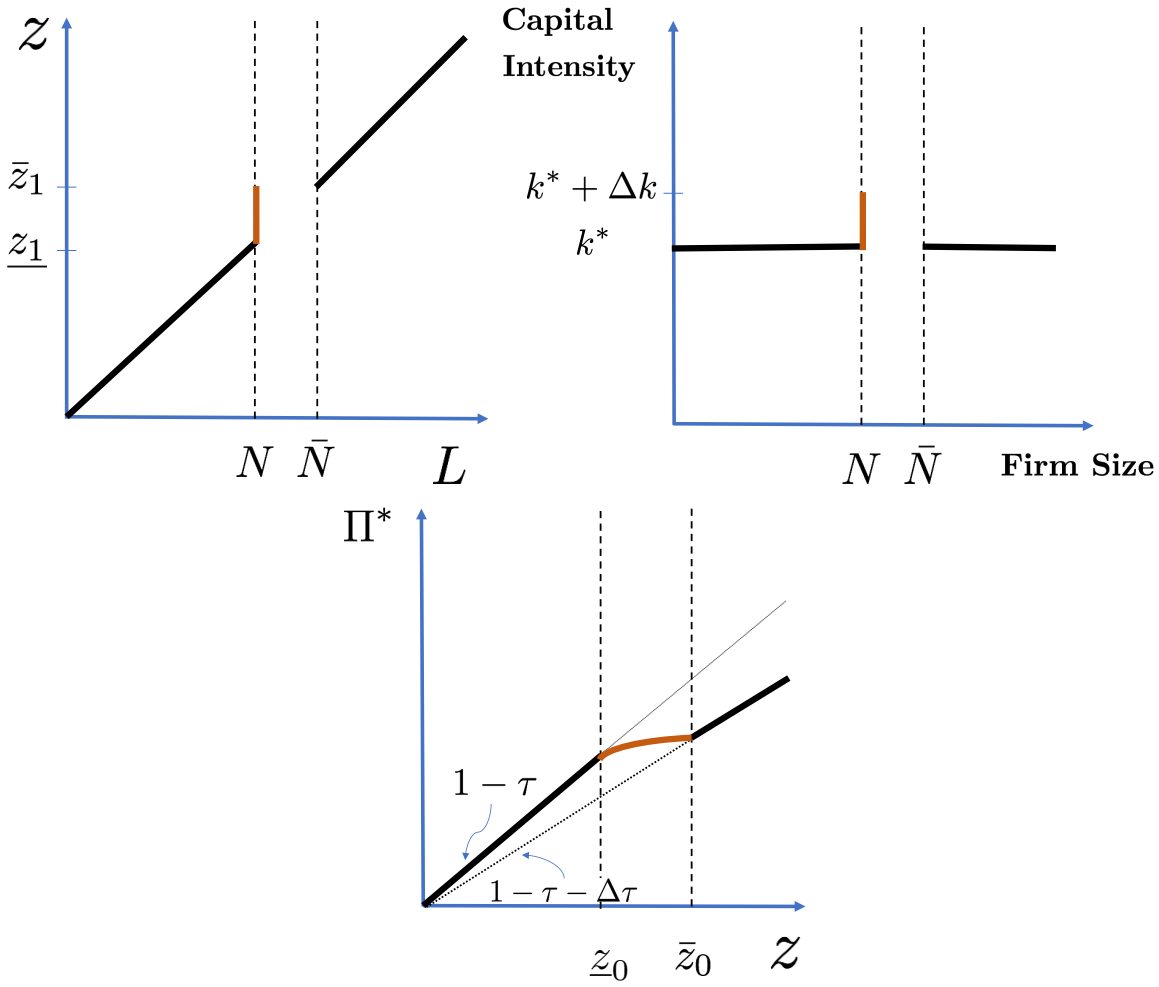
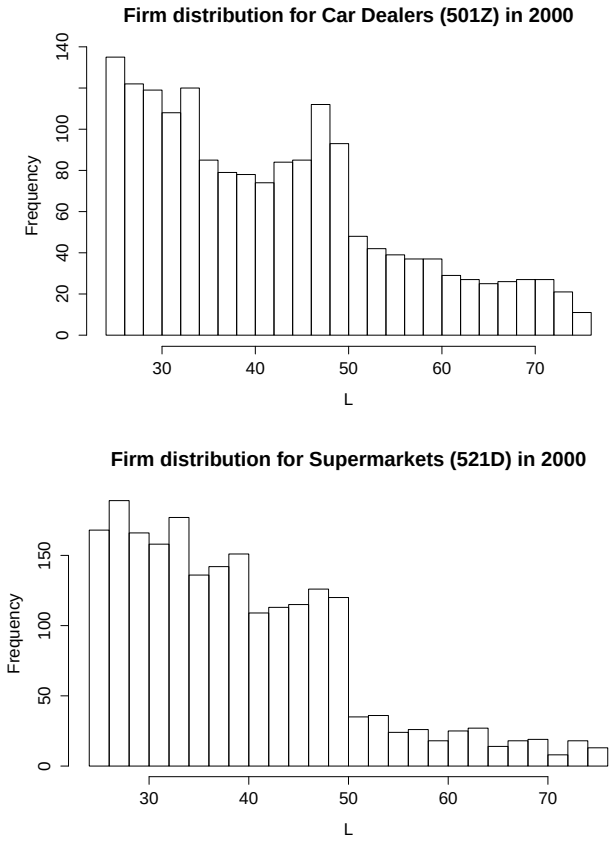


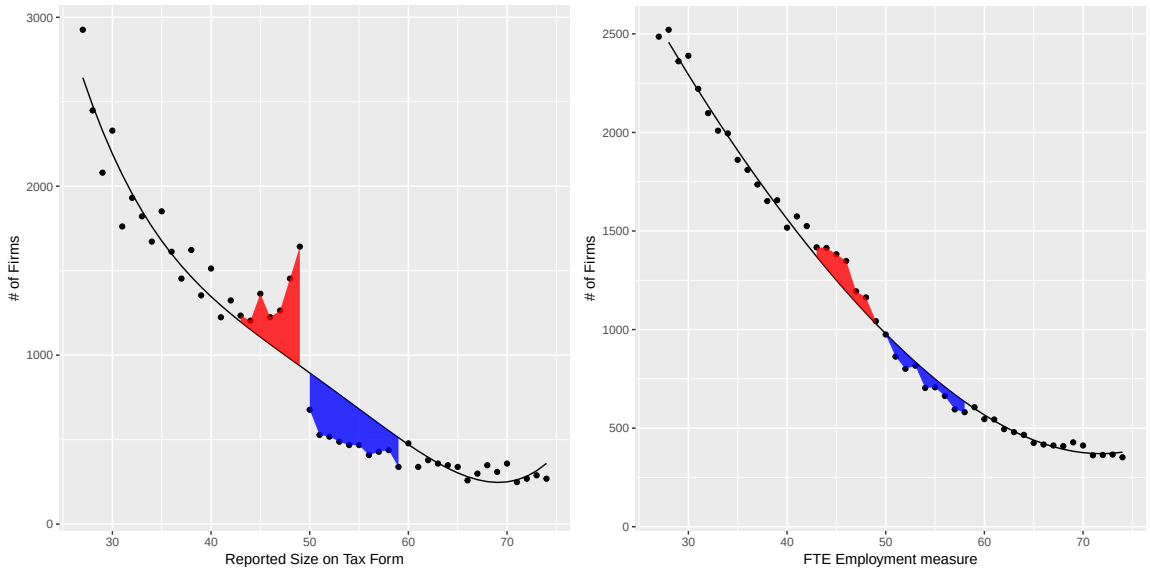
Figure 1.4: Firm Size distributions at selected 4-digit sectors



Source: FICUS

Figure 1.5: Excess Mass in the Two Datasets

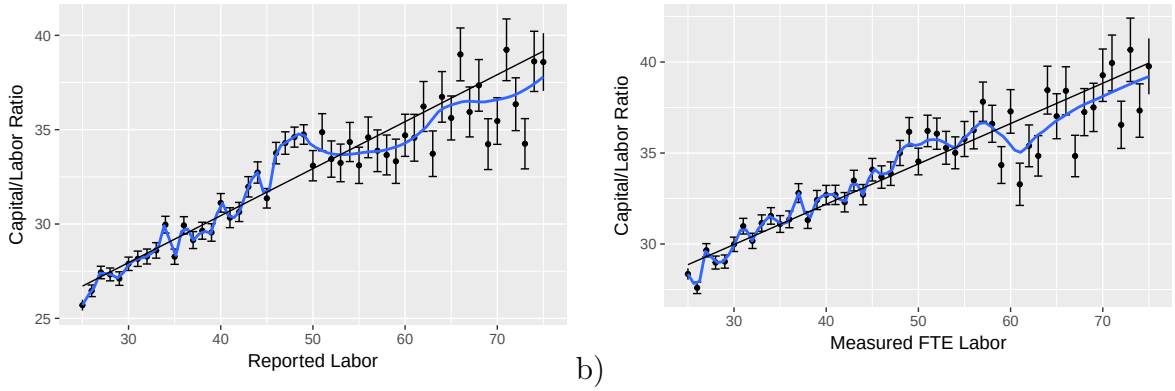
a) At the 50-employee threshold



b) At the 10-employee

The excess mass are computed in each dataset by fitting a 6th order polynomial The left panel uses the self-reported data while the right panel uses the measured data. Source: FICUS and DADS

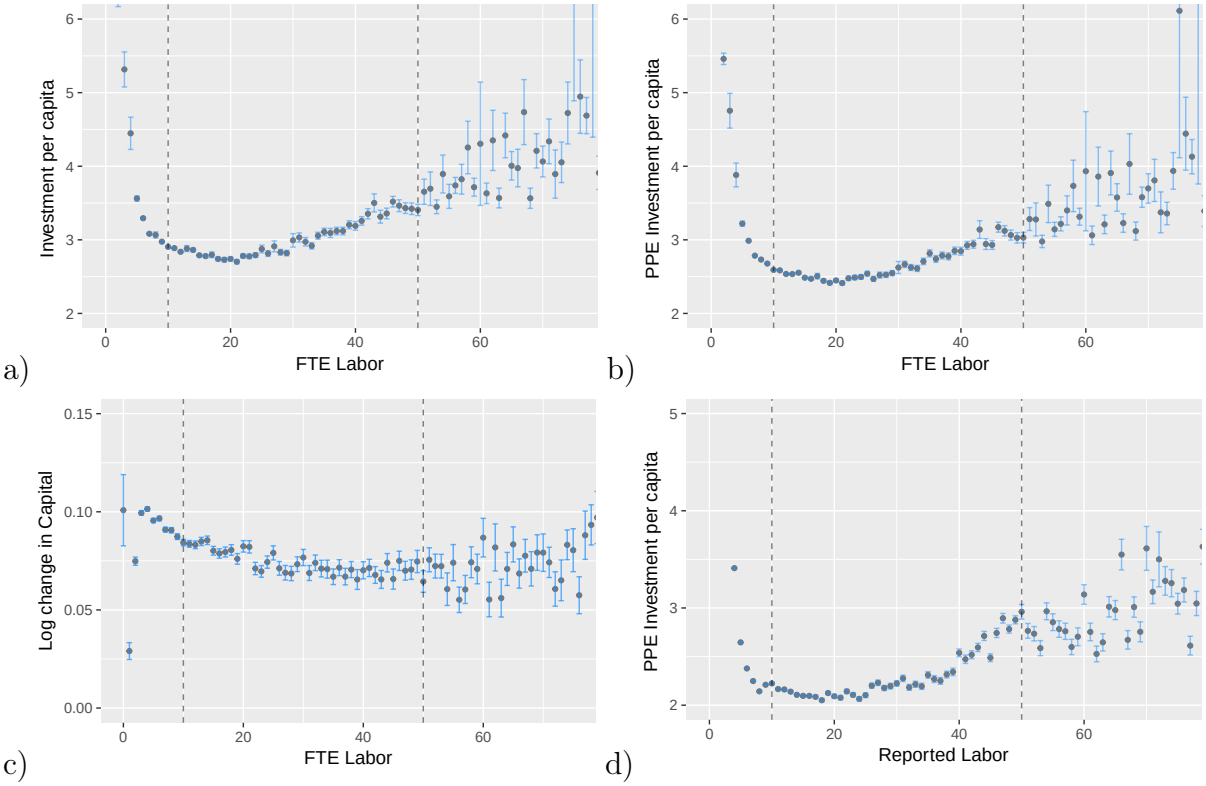
Figure 1.6: Distortions in Capital-Labor ratios



a) b)

Note: Panel a) shows the distortion in capital-labor ratio where the reported labor variable is used. Panel b) shows the capital-labor for the same firms but using the FTE measure computed by the administration instead. Source: FICUS and DADS.

Figure 1.7: Distortions in Investment per capita



Note: I report the investment response around the threshold for three measures of investment: a) total gross investment, b) gross investment in property, plant, and equipment, and c) the log change in PPE capital. In all panels except for d) the x-axis is the FTE measured Labor. Only when the x axis is the self-reported labor does investment seem abnormally high before the cutoff.

Figure 1.8: Notch with Frictions

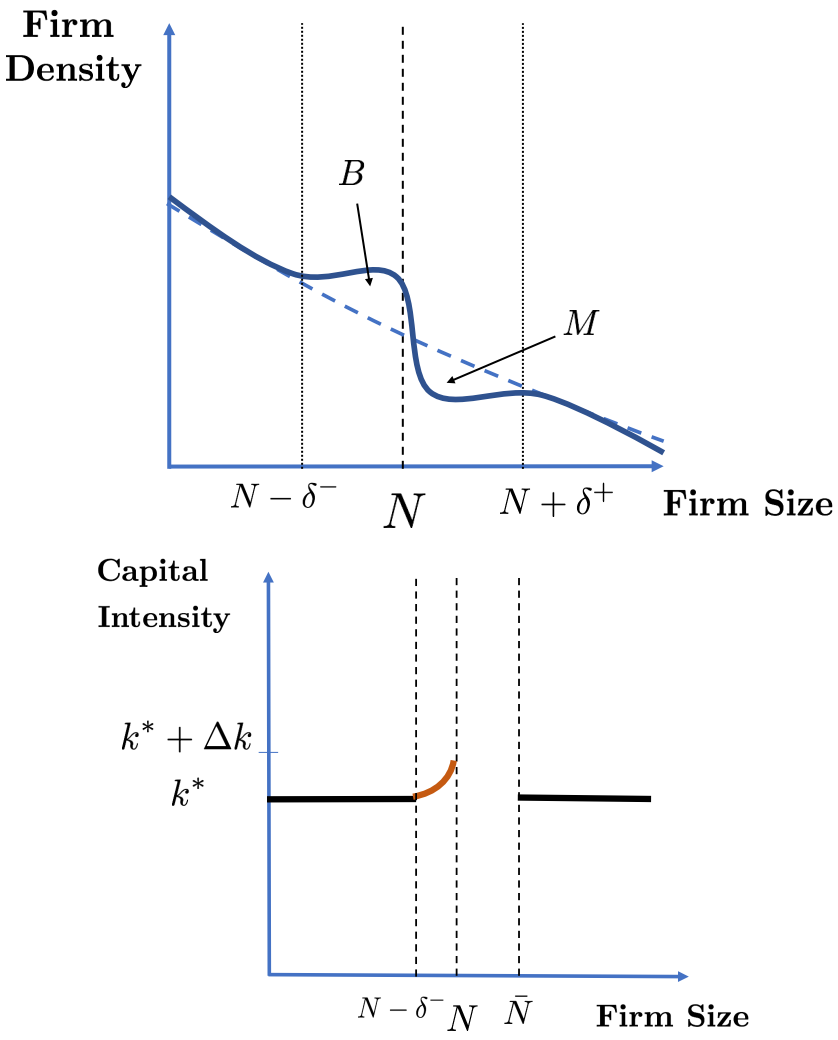
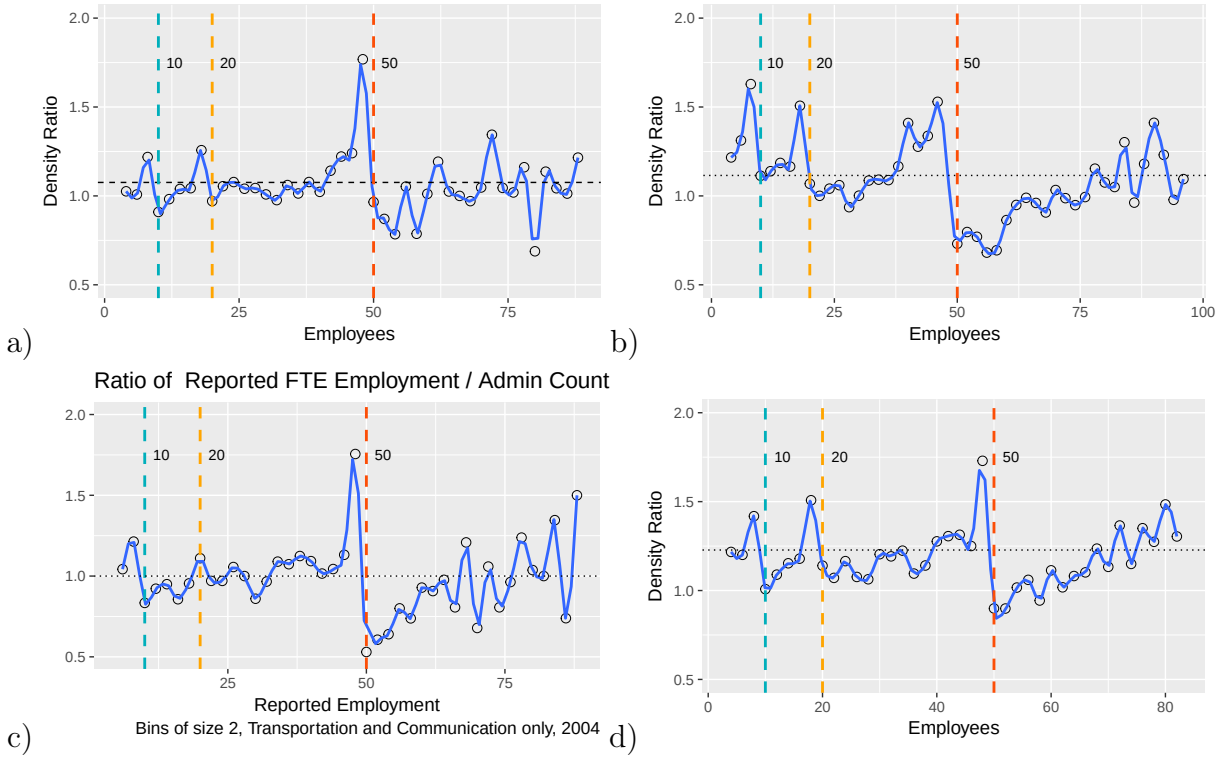
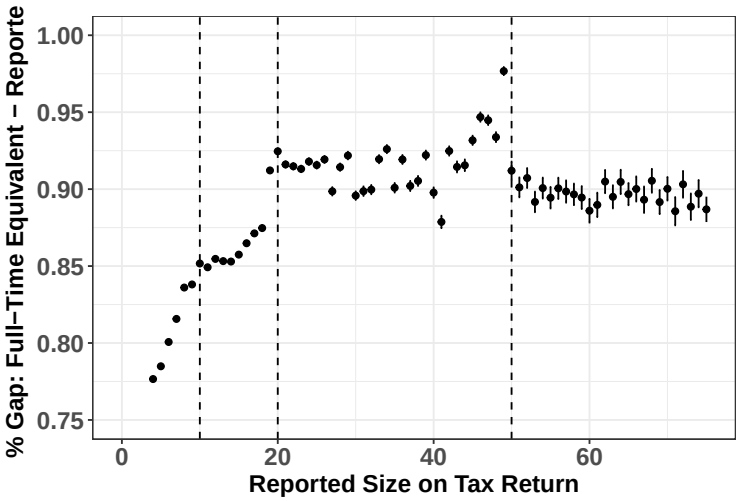


Figure 1.9: Misreporting Evidence: Ratios of Firm-Size Densities at the Sector Level



Note: the firm size densities are computed for each sector using alternatively reported and FTE-measured employment. Then I compute the ratio of these two densities (y axis). The plots shows these density ratios for a) the manufacturing sector, b) the retail sector, c) the transportation and communication sector, d) Business services. For instance, there are 75% more firms with 49 reported employees than with 49 FTE-measured employees. The dotted line is the ratios' average, excluding the bunching regions. Source: FICUS 2004-2007

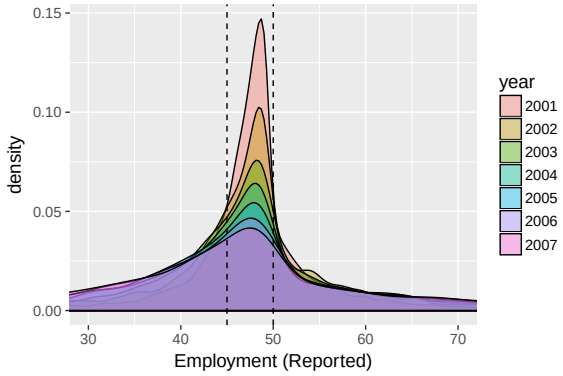
Figure 1.10: Misreporting Evidence: Self-Reported vs Administration's Full-Time Employment measure.



b)

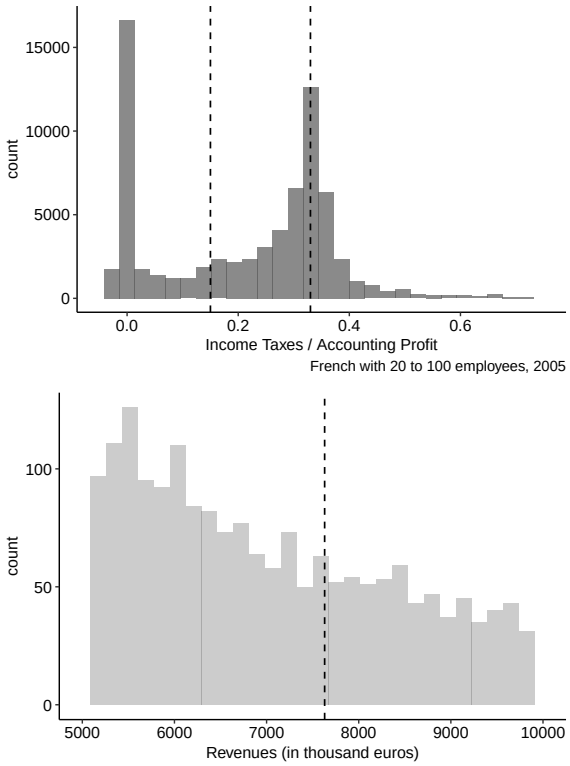
Note: the figure plots the percentage gap between the Social Security Labor measure and the self-reported labor measure. The formula used by the administration to compute the FTE measure is top-censored at 1, which creates a downward bias of 10% on average. Because o f misreporting this gap shrinks abruptly at the 20- and 50-employee threshold. Under 50-employee, the standard errors are tiny and hardly distinguishable from the point estimate. Source: Autor's calculation from FICUS and DADS.

Figure 1.11: Bunching Persistence around the 50-employee threshold



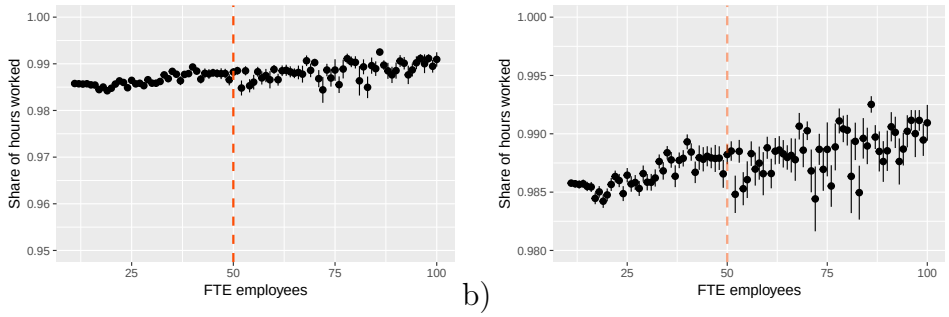
Note: This is the employment density over time of the firms who were bunching at the 50-employee threshold in 2000. About half of them are bunching the next period as well and 20% of them are still bunching 7 years later. Source: Author's calculation from FICUS.

Figure 1.12: Effective tax rates



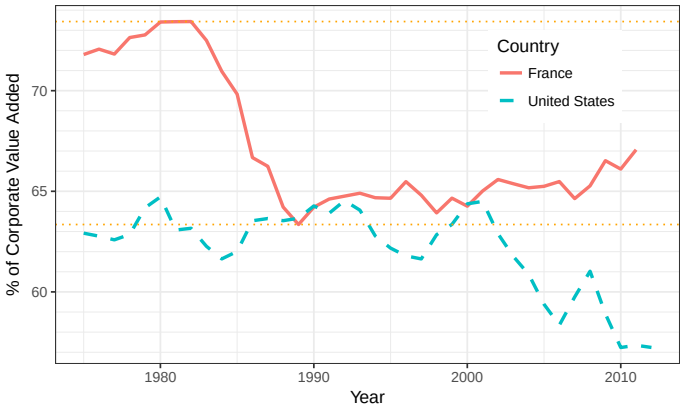
Note: The top panel shows the average effective income tax rate paid by firms with 20 to 100 employees. It implies that for almost all of these firms turning profits, the relevant marginal corporate income tax rate is the standard rate at $33, 1/3\%$. The bottom panel shows that there is no significant bunching at the kink point between the reduced and standard rates. Source: Author's calculation from FICUS.

Figure 1.13: Primary jobs account for 99% of the hours worked



Note: these figures show that the Full-Time Employment measure in the Social Security Data is a very reliable measure of employment at the firm level. It is unproblematic to leave aside "secondary jobs" as they account for less than 1% of the hours worked, and this fraction is stable with firm size. For each firm, I use the matched employer-employee data to sum up the total number of hours worked in primary jobs and in secondary jobs down. I then compute the share of hours corresponding to primary jobs. Finally I bin firms per FTE size (rounded) and compute the mean and standard error. Source: DADS 2004

Figure 1.14: Evolution of the Labor Shares in France and in the United States



The corporate labor share is the share of labor compensation in the value added of the corporate sector, that is, excluding the government sector, sole proprietors and unincorporated enterprises. This concept arguably makes cross-country comparisons more transparent (Karabarbounis and Neiman 2014). Source: author's calculations from OECD data.

Table 1.1: Descriptive statistics: Comparing the Firm Size Distributions

	FICUS	DADS (FTE over whole year)	DADS (FTE on 12/31)
# of firms with	2,504,173	1,573,610	1,573,594
0 employees	1,215,064	123,673	236,816
> 0 employees	1,289,109	1,449,937	1,336,778
1	210,564	245,395	224,118
9	36,263	28,289	28,449
10	18,098	21,837	21,890
18	12,273	7,276	7,434
19	11,783	6,540	6,623
20	9,661	5,974	5,870
47	2,547	1,292	1,304
48	3,021	1,222	1,225
49	1,647	1,047	1,097
50	678	1,018	1,012

Note: The difference in the total number of firms is mainly due to the firms with no employee as the DADS (Social Security data) registers payroll data. Starting from ... FICUS records income of partnerships, medical practices, entrepreneurs etc. Source: FICUS and DADS 2004 .

Table 1.2: Capital-Labor Distortions: Reported Employment

	Dependent variable:					
	log(Capital-Intensity)					
	OLS					*Sector FE*
	(1)	(2)	(3)	(4)	(5)	(6)
L	0.005***	0.008***	0.004***	0.005***		0.006***
	(0.0001)	(0.0001)	(0.0001)	(0.0001)		(0.0001)
L^2		−0.00003***				
		(0.00000)				
Bunching			0.184***	0.095***	0.083***	0.088***
			(0.009)	(0.028)	(0.028)	(0.026)
L<45				−0.058**	−0.274***	−0.005
				(0.026)	(0.026)	(0.025)
L>55				−0.145***	0.035	−0.125***
				(0.027)	(0.027)	(0.025)
Constant	2.538***	2.518***	2.539***	2.591***	2.853***	
	(0.001)	(0.001)	(0.001)	(0.027)	(0.026)	
Obs.	2,360,558	2,360,558	2,360,558	2,360,558	2,360,558	2,340,215
R^2	0.004	0.004	0.004	0.004	0.002	0.115
Adj. R^2	0.004	0.004	0.004	0.004	0.002	0.115
Res. S.E.	1.286	1.286	1.286	1.286	1.287	1.207
F Stat.	8,591.332***	4,629.088***	4,501.677***	2,281.012***	1,930.439***	

*p<0.1; **p<0.05; ***p<0.01

Table 1.3: Capital-Labor Distortions: FTE-measured Employment

	Dependent variable:						
	log(Capital-Intensity)						
	OLS					*Sector FE*	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
etpr	0.004***	0.004***	0.004***	0.001***			
	(0.0001)	(0.0001)	(0.0001)	(0.0001)			
L						0.003***	0.001***
						(0.0001)	(0.0001)
bunching			0.147***	0.024	0.022	0.051**	0.040*
			(0.009)	(0.027)	(0.027)	(0.026)	(0.024)
L<45				−0.221***	−0.268***	−0.129***	−0.133***
				(0.026)	(0.026)	(0.025)	(0.023)
L>55				0.054**	0.075***	−0.073***	−0.049**
				(0.026)	(0.026)	(0.025)	(0.023)
Constant	2.714***	2.714***	2.716***	2.955***	3.013***		
	(0.001)	(0.001)	(0.001)	(0.026)	(0.026)		
Obs.	1,957,932	1,957,932	1,957,932	1,957,932	1,957,932	1,957,932	1,957,932
R^2	0.002	0.002	0.002	0.003	0.003	0.120	0.228
Adj. R^2	0.002	0.002	0.002	0.003	0.003	0.120	0.228
Res. S.E.	1.248	1.248	1.247	1.247	1.247	1.181	1.106
F Stat.	3,673.795***	3,673.795***	1,970.334***	1,306.335***	1,692.167***		

*p<0.1; **p<0.05; ***p<0.01

Table 1.4: Dispersion in Income Shares and Elasticity by Sectors

NAF code	Sector	Dispersion (χ_{ni})	Firm-level Elasticity (σ_i)	Demand Elasticity (Benchmark)	Sectoral Elasticity
70	Real Estate	0.77	0.12	1.3	0.26
74	Business Services	0.37	0.14	1.3	0.19
52	Retail	0.10	0.04	1.3	0.06
55	Hotel and restaurants	0.18	0.06	1.3	0.10
45	Construction	0.19	0.11	1.3	0.15
50	Wholesale and Trade	0.07	0.21	1.3	0.22
24	Chemical Industry	0.27	0.18	1.3	0.21

Note: Elasticities measured using the 50-employee threshold. The dispersion coefficient is measured using net assets. NAF is the acronym of the French classification of economic activity. Activities are ranked in decreasing order of the amount of capital they command. Source: FICUS 2004.

Table 1.5: Sectoral Dispersion in Income Shares and Elasticities

NAF code	Sector	Firm-level Elasticity σ	Share α_n
74	Business Services	0.19	0.37
52	Retail	0.06	0.10
55	Hotel and restaurants	0.10	0.18
45	Construction	0.15	0.19
50	Wholesale and Trade	0.22	0.07
15-35	Manufacturing	0.12	0.21

Note: elasticities are measured at the 50-employee threshold. Dispersion is measured using net assets. NAF is the acronym of the French classification of economic activity. Activities are ranked in decreasing order of the amount of capital they command. Source: FICUS 2004.

Table 1.6: Structural Estimates

	Parameter	Value	S.E
$\hat{b}$	Bunching Mass	0.09%	0.002
$\Delta\hat{k}$	Capital distortion	0.011	0.017
τ	Regulation Cost	0.12 %	0.0023
σ	Elasticity of Substitution	0.18	0.13
θ	Variance of the Optimization Error	0.05	0.17
γ	Convexity of misreporting cost	3	.
ν	Returns to Scale	0.85	.

Note: Firms in manufacturing Sector (NAF 15-35) 50-employee threshold. Source: Ficus 2004-7.

CHAPTER 2

Macroeconomic Effects of Competition Distortions

This paper builds a quantitative framework to assess the macroeconomic implications of the distortions to market structure created by cartels and by anti-competitive behavior. We analyze these distortions in an extension of the oligopolistic framework of Atkeson and Burstein (2008) which features heterogenous firms and endogenous markups. We find that the direct negative welfare impact of cartels is compounded by an umbrella pricing effect, whereby firms outside of the cartel also raise their prices. We test the predictions of the model by combining exhaustive administrative firm micro-data with a novel dataset on firm- and sector-level collusion cases constructed from textual analysis of the French Competition Authority's decisions.[1]

2.1 Introduction

What are the effects of competition distortions on resource allocation and aggregate welfare? How do competitors respond to cartels? How much scope is there for competition policy to improve on macroeconomic outcomes? The answers to these questions are important but have been elusive. In an influential article, **harberger1954monopoly** suggested that the inefficiency costs generated by monopolies in the U.S. cannot exceed a tenth of a percent. Harberger used a simple calibration and aggregate data to estimate the sizes of the triangles that now bear his name. In contrast, we take a more direct route and focus on those cartels that have been investigated. In practice, we analyze two decades of antitrust decisions taken by the French Competition Authority, and examine their impacts at the firm and sector

[1] joint work with Ludovic Panon

level, as well as on the aggregate French economy. Our methodology uses textual analysis tools to build a detailed database on collusive cases from the French Competition Authority decisions. We then combine these data with exhaustive administrative firm-level data. To understand the aggregate impact of collusive behaviors on the whole economy, we develop a new framework that features heterogenous firms, endogenous markups, and the possibility to form cartels.

Our tractable oligopolistic model builds on Atkeson and Burstein 2008. A finite number of firms compete with each other à la Cournot and facing a Constant Elasticity of Substitution demand.[2] This model generates variable markups as a firm's demand elasticity is a function of its market share: firms with a relatively high market share face less elastic demand which allows them to charge higher markups. We extend this framework to allow firms to form cartels. To do so, we embed a cross-ownership model of firms à la Azar, Schmalz, and Tecu 2018 into the Atkeson and Burstein 2008 framework. In our model, a cartel member maximizes its profits by internalizing the effect of its decision (quantity or prices) on other cartel members. A key advantage of our framework is that the effects of collusion on firm-level outcomes and welfare can be derived analytically and easily simulated. We provide formulas to estimate at the first order the predicted price impact of any cartel given the equilibrium market shares.

The model predicts that *both* colluding and non-colluding firms increase their markups and prices. In particular, we uncover an "umbrella pricing" channel whereby firms who are *not* part of the cartel find it to raise their prices as well following the formation of the cartel. The magnitudes of these effects depend on the relative market shares of the competitors. Because prices increase across-the-board, the price index rises and the output of the sector in which the cartel operates decreases. We test these empirical predictions using our newly created firm-level database on anti-competitive conduct.

Our novel cartel dataset uses all the information contained in the sentencing decision. This allows us to measure precisely the members of the cartels, its duration, the severity of

[2]We consider a variant of the model with price competition à la Bertrand is detailed in the Appendix, with similar qualitative implicatioms.

the collusion and the fines handed down to each firm. We the merge our firm-level cartel database with exhaustive balance sheet data. This administrative dataset covers the universe of French firms allows us to measure the impact of anti-competitive practices not only on the firms targeted by the investigation, for which some information is already contained in the Competition Authority's decisions, but also on all their competitors.

Our preliminary findings focus on measuring the effects of collusive behaviors on markups, market shares and employment, controlling for various firms characteristics, including their revenue productivities. First, most firms charge markups close to the monopolistic competition benchmark of constant markups, because they have trivial market shares and hence are unable to exploit their market power. However, the largest firms that account for up to around 25 percent. This is consistent both with the model and with findings by Hottman, Redding, and Weinstein 2016 for U.S. firms. Second, we find that firms operating in sectors where at least one firm is behaving anti-competitively lose market shares but that the effect is in fact positive for the firms that collude. This suggests that the higher markups charged by colluders are not competed out and points to potentially large inefficiencies. However, because larger firms are more likely to join a cartel, our preliminary estimates are not immune to sample selection bias. In ongoing work, we develop synthetic control methods to correct for this source of endogeneity and identify a causal effect of collusion on firm-level outcomes.

The paper proceeds as follows. In Section 2, reviews the related literature. In Section 3, introduces our model. Section 4 provides a detailed description of the data used in Section 5's empirical framework. Our preliminary results are presented in Section 6. Section 7 concludes.

2.2 Related Literature

In this paper, we embed a theory of collusive behavior in an otherwise standard macroeconomic model with heterogeneous firms. In what follows, we review the key contributions on which this paper builds.

Market Structure in Macroeconomics. Market structure is the object of a renewed interest in macroeconomics.[3] Recent work connects increasing concentration in markets with the fall of the labor share Autor et al. 2017 or the increase in markups De Loecker and Eeckhout 2017. A burgeoning literature also examines the monoposony effects of mergers Azar, Marinescu, and Steinbaum 2017; Azar, Schmalz, and Tecu 2018. In our setting, the labor market is perfectly competitive and we abstract from these welfare effects. Finally, Edmond, Midrigan, and Xu 2018 assess the role of markups in generating potentially large distortions in the economy. Our paper, on the other hand, focuses on a specific type of distortion arising from anti-competitive practices encompassing collusions and abuse of dominant positions. We can therefore precisely estimate the source of welfare loss arising from increases in concentration and/or markups.

Theory of Cartels. The theoretical literature on cartels is extremely rich and we do not claim to provide an exhaustive list of works. Some of these studies are summarized in Tirole 1988. The seminal contribution of Stigler 1964 disseminated the idea that cartels are by their very nature unstable. Indeed, cheating incentives are strong and undermine the existence and stability of collusive cases. Much more recent is the work by Bos and Harrington 2010 who show that larger firms might have a strong incentive to form a cartel with smaller firms being able to increase their prices as the larger firms' prices serve as an umbrella.[4] Compared to these studies, our theoretical model is more tractable as we can easily estimate and simulate it, and derive empirical predictions to take to the data.

Empirical Analysis of Cartels. The empirical study of cartels and their impact on welfare is limited by the fact that secret agreements are by definition hard to observe. However, it is possible to focus on specific cartels operating in particular industries. Levenstein and Suslow 2006 summarize findings in the literature: the picture that emerges is not that predicted by Stigler 1964. Indeed, on average, cartels are not short-lived and antitrust activity

[3]The relevance of market structure for macroeconomic policy was the central topic of the 2018 Jackson Hole's conference, entitled "Changing Market Structure and Implications for Monetary Policy".

[4]Another important paper is Selten 1973's work on the optimal number of cartel members.

is a likely cause for cartel death **levenstein2011breaking**. In an interesting study, Symeonidis 2008 finds that *manufacturing industries* that were cartelized experienced slower labor productivity growth than those that were not. Our data is more disaggregated as we match information on anti-competitive firms to firm-level balance-sheet data. This allows us to assess the causal effect of collusive practices on firm-level outcomes.

2.3 Model

We develop a model in which heterogeneous firms choose their markups endogenously along the lines of Atkeson and Burstein 2008. The model allows for Cournot and Bertrand competition. The economy is made of a continuum of sectors, but in each sector, only a *finite* number of firms compete. The firms are therefore "large in their own sectors, but small in the economy as a whole" Neary 2003. In equilibrium, firms' markups are proportional to their market share. We extend this framework to allow groups of firms to form cartels. Collusion affects how firms take into account the impact of their production and pricing decisions on the sectoral output and price level. Colluding in our model is akin to cross-ownership, and produces similar competition distortions (Azar, Schmalz, and Tecu 2018). Collusion unambiguously raises prices and is harmful to consumers.[5]

2.3.1 Oligopolistic Competition

We keep the demand side of the economy voluntarily stark in order to focus on the supply implications of the competition shocks. All the important economic decisions are made by the firms. An infinite-lived representative household maximizes a time-separable utility

$$\mathbb{E}\sum_{i=0}^{\infty}\beta^t \log\left[c_t^{\delta}\left(1-l_t\right)^{1-\delta}\right]$$

[5]In the text, we solve the model under Cournot competition. See the Appendix for the version with Bertrand Competition.

The first order conditions for the household are standard and yield the familiar intra-temporal tradeoff between consumption and leisure:

$$\frac{1-\delta}{\delta}\frac{c_t}{1-l_t} = \frac{W_t}{P_t} \tag{2.1}$$

For simplicity we drop the time subscript and focus on the stationary case. The production side of the economy consists in a continuum of sectors indexed by $i \in [0,1]$. Final consumption c is produced by a competitive firm that combines the outputs from all the sectors y_i with a CES technology with demand elasticity η:

$$c = \left[\int_0^1 y_i^{\frac{\eta-1}{\eta}} di\right]^{\frac{\eta}{\eta-1}} \tag{2.2}$$

The inverse demand function for each intermediate output from sector i is given by:

$$\frac{P_i}{P} = \left(\frac{y_i}{c}\right)^{-\frac{1}{\eta}} \tag{2.3}$$

where P, the price index for final consumption representing the "true cost of living", is an harmonic mean of the sectoral prices:

$$P = \left[\int_0^1 P_i^{1-\eta} di\right]^{\frac{1}{1-\eta}} \tag{2.4}$$

Each sector is populated by a finite number of firms K_i. Because each firm has a non-zero measure, its decisions have an impact on its competitors' decisions. Firms are "large in the small but small in the large" (Neary 2003) *i.e.* they are "small" with respect to the economy but "big" in their sector. The output in sector i is a composite of the firms' outputs, combined with a CES technology with elasticity parameter ρ:

$$y_i = \left[\sum_{k=1}^{K_i} (q_{ik})^{\frac{\rho-1}{\rho}}\right]^{\frac{\rho}{\rho-1}} \tag{2.5}$$

The price index in sector i is given by

$$P_i = \left[\sum_{k=1}^{K_i} (P_{ik})^{1-\rho}\right]^{\frac{1}{1-\rho}} \tag{2.6}$$

We make the following assumptions:

Assumption 1. *Goods are imperfect substitutes* $\rho < \infty$.

Assumption 2. *Goods are more substituable within than between sectors* $1 < \eta < \rho$.

Assumption 3. *Firms play a static game of quantity competition (Cournot).*

Assumption 4. *The Productivity distribution is lognormal* $\log z_{ik} \sim \mathcal{N}(0, \theta)$.

Several remarks are in order. Assumption 1 is standard. Assumption 2 is crucial for our analysis to go through. It guarantees that firms' markups are increasing in market shares (see equation (2.10)) . Assumption 3 can be replaced by Bertrand competition which does not alter our qualitative results. Assumption 4 is made in order to keep the quantitative exercise simple but is unimportant for the proof. We now solve the model. The *vectors* of prices $\mathbf{P}$ and quantities $\mathbf{q}$ maximize profits

$$\max_{P,q} \mathbf{P}.\mathbf{q} - \mathbf{q}.W.\left[\frac{1}{z_{ik}}\right]_{(i,k)} \tag{2.7}$$

subject to the inverse demand functions

$$\left(\frac{\mathbf{P}}{P}\right) = \left(\frac{\mathbf{q}}{y_i}\right)^{-\frac{1}{\rho}} \left(\frac{y_i}{c}\right)^{-\frac{1}{\eta}} \tag{2.8}$$

When the firms compete *à la Cournot*, the first order conditions imply that the equilibrium price is a markup over the marginal production cost

$$P_{ik} = \frac{\epsilon(s_{ik})}{\epsilon(s_{ik}) - 1} \frac{W}{z_{ik}} \tag{2.9}$$

where the firm-specific inverse elasticity is a linear combination of the within- and between-sector elasticities

$$\epsilon(s_{ik}) = \left[\frac{1}{\rho}(1 - s_{ik}) + \frac{1}{\eta} s_{ik}\right]^{-1} \tag{2.10}$$

where

$$s_{ik} = \frac{P_{ik} q_{ik}}{\sum_{j=1}^{K} P_{ij} q_{ij}} \tag{2.11}$$

is the market share of firm k in its sector i. Hence firm-level markups are heterogenous and endogenous, reflecting their comparative advantage within their sector. The attractive feature of equation (2.10) is that while the underlying productivity is not observed, the

market shares are. Moreover, we can further eliminate the quantities q_{ik} and obtain a system only in terms of prices P_{ik}. To see this, from the CES assumption, the market shares can be expressed solely in terms of prices:

$$s_{ik} = \frac{(P_{ik})^{1-\rho}}{\sum_{j=1}^{K}(P_{ij})^{1-\rho}} \tag{2.12}$$

as quantities produced are related to prices from the inverse demand functions (2.3) and (2.8)

$$\begin{aligned} \frac{P_{ik}}{P} &= \left(\frac{q_{ik}}{y_i}\right)^{-\frac{1}{\rho}}\left(\frac{y_i}{c}\right)^{-\frac{1}{\eta}} \\ &= \left(\frac{q_{ik}}{y_i}\right)^{-\frac{1}{\rho}}\left(\frac{P_i}{P}\right) \end{aligned} \tag{2.13}$$

Another way of characterizing the equilibrium markups is to express them in terms of the prices selected by the firms. Firm k's markup is a harmonic mean of the two markups associated with the CES demand, $\mu_\rho \equiv \frac{\rho}{\rho-1}$ and $\mu_\eta \equiv \frac{\eta}{\eta-1}$, where the weights are given by the market share, or, equivalently, by the firm's price relative to the sectoral price index

$$\frac{1}{\mu(P_{ik})} = \left[\frac{1}{\mu_\rho} + \left(\frac{1}{\mu_\eta} - \frac{1}{\mu_\rho}\right)\left(\frac{P_{ik}}{P_i}\right)^{1-\rho}\right] \tag{2.14}$$

Small firms care mostly about the competition coming from firms in the same sector whereas larger firms dominant in their sector internalize some of the substitution effect between sectors. As the market share converges to 1, the relative price $\frac{P_{ik}}{P_i}$ converges to 1 from above and the markup tends to μ_η, the markup associated with the between-sector CES demand. For quantitative explorations of the models, we follow the baseline calibration in Atkeson and Burstein 2008 and choose a within sector elasticity of 10. That is, most of the market power is coming from the lack of substitution between sectors. We test the robustness of our results to different values for the elasticity. See the appendix for numerical explorations with other values for these elasticities.

General equilibrium. The model's equilibrium is found by solving a fixed point problem in the aggregate variables $\{\mathbf{P}, W, c, l\}$. The solution procedure takes three steps: i) given $\mathbf{P}, c, W$ we solve for the prices and quantities in every sector; then ii) using, in each sector,

the system of N_i nonlinear equations (2.9) by substituting the expression for the market share (2.12) ; finally iii) we check that the household's first order conditions (2.1) holds.[6]

Pass-through To understand the impact of productivity shocks we can characterize the pass-through effects at the first-order when the economy is initially at the equilibrium. Let us first define, Ω, the matrix that captures the cross-price elasticities.

$$\mathbf{\Omega}_{kj} = \begin{cases} -\mu_k \left(\frac{1}{\eta} - \frac{1}{\rho}\right)(\rho - 1)\, s_k \,(1 - s_k) & j = k \\ \mu_k \left(\frac{1}{\eta} - \frac{1}{\rho}\right)(\rho - 1)\, s_k \cdot s_j & j \neq k \end{cases} \tag{2.15}$$

The matrix Ω captures the interdependence between the firms' optimal pricing decisions around the equilibrium. Given the CES demand system, this effect is channelled through the sectoral price index and firms with larger market shares create larger responses. The diagonal terms captures the fact that faced with an exogenous positive shock to its price (say, a increase in marginal cost), the firm responds by lowering its price. In contrast, notice that all the off-diagonal terms are positive, attesting that the presence of *pricing complementarities* between the firms. We will see in the next section that the pricing complementarities are enhanced in the presence of collusion.

Proposition 6. *At the first order, the price impact of a productivity shock is:*

$$d \ln \mathbf{P} = -T d \ln \mathbf{z}$$

where $T \equiv [I - \mathbf{\Omega}]^{-1}$*, and the pass-through is imperfect, in particular:*

$$0 \geq \rho(T) < 1$$

Proposition 7. *1) The umbrella pricing effect is larger for larger firms. 2)More concentrated markets have a lower pass-through.*

[6]Entry would create more complications. It would require, in order to pin down the equilibrium, to specify the order in which potential entrants with different productivity draws enter the market.

Aggregate Productivity and Markups. The quantity of final output can be represented by an aggregate production function $Y = A \cdot L$. As shown in Edmond, Midrigan, and Xu 2015, the first-order condition for the optimal use of labor and the labor market clearing condition yield:

$$A = \left[\int_0^1 \left(\sum_{k=1}^{K_i} \frac{1}{z_{ik}} \frac{y_{ik}}{Y} \right) di \right]^{-1}$$

where aggregate productivity A is a quantity-weighted harmonic mean of firm productivities. The aggregate markup in the economy, defined as the ratio of the aggregate price over marginal cost $\mu_{agg} = \frac{P}{W/A}$, can similarly be expressed as a revenue-weighted harmonic mean of firm productivities

$$\mu_{agg} = \left[\int_0^1 \left(\sum_{k=1}^{K_i} \frac{1}{\mu_{ik}} \frac{p_{ik} y_{ik}}{PY} \right) di \right]^{-1}$$

Alternatively, the aggregate productivity can be written in terms of the firm productivities and their relative markups

$$A = \left[\int_0^1 \left(\frac{\mu_k}{\mu_{agg}} \right)^{-\eta} z_k^{\eta-1} di \right]^{\frac{1}{\eta-1}}$$

where z_k is the sector-level productivity given by

$$z_k = \left[\sum_{k=1}^{K_i} \left(\frac{\mu_{ik}}{\mu_k} \right)^{-\rho} z_{ik}^{\rho-1} \right]^{\frac{1}{\rho-1}}$$

and $\mu_k = \frac{P_k}{W/z_k}$ is the sectoral markup. An increase in markup dispersion therefore reduces the allocative efficiency of the economy.

2.3.2 Collusion as Cross-ownership

We now analyze the effects of collusion in this economy. In each industry, firms can form a cartel $\mathcal{C}$ of any size. When firms collude, their productivities are unchanged but their pricing decisions are altered because they partially internalize the effects of their decisions on the other members of the cartel. As a consequence the markups of the colluding firms rise and so does the price index of their sector. Collusion therefore harms consumers.

Formally, we model collusion as an implicit joint venture between cartel members. The distortions in the firms' incentives operate just like the distortions created by common ownership claims between firms Azar, Schmalz, and Tecu 2018; O'brien and Salop 1999. [7]Instead of simply maximizing its own profits, each cartel member internalizes part of its price impact and maximizes instead a linear combination of all the cartel members' profits, just like one would expect in the presence of cross-ownership patterns. This flexible formulation allows analytical derivations of collusion of various intensities and sizes, in a manner that is much more tractable, elegant, and transparent that comparative statics between oligopolistic outcomes of pricing games between a dozen firms or so.

Consider an industry with K firms and denote by Π_k the profit function of firm k. Theories of the firm [8] usually make a distinction between financial ownership – a claim to a share of profits – and corporate control – the right to participate in the firm's production decisions –. Our model can flexibly accomodate the two. Let β_{lj} denote the share of firm j which is owned by investor l and γ_{lj} investor l's controlling share of firm j. The profits of investor l correspond to the portfolio $\pi^l = \sum_k \beta_{lk}\Pi_k$. The managers of firm k maximize a weighted average of the firm's shareholders portfolios, where the weights depend on the controlling shares

$$\tilde{\Pi}_k = \sum_l \gamma_{lk} \sum_j \beta_{lj}\Pi_j \tag{2.16}$$

After rearranging and dividing by $\sum_l \gamma_{lk}\beta_{lk}$ we obtain the distorted profit function of firm k in a collusive environment[9]

$$\tilde{\Pi}_k = \Pi_k + \sum_{j\neq k} \frac{\sum_l \gamma_{lk}\beta_{lj}}{\sum_l \gamma_{lk}\beta_{lk}}\Pi_j \tag{2.17}$$

The sum on the right hand side is a linear combination of other firms' profits. Suppose that there are $M_{\mathcal{C}}$ firms involved in the cartel. One can view the cartel as a joint-venture between these firms, whereby each retains full control of its production ($\gamma_{kk} = 1$) but cedes a share

[7]See also Azar and Vives 2018 for a model where firms are also large in the economy.

[8]see Shleifer and Vishny 1997's survey of Corporate Governance

[9]The rescaling is innocuous regarding the incentives once in the cartel but would matter for the decision to become a member of the cartel or not.

of its profits to the other members of the cartel.[10] If each firm keeps a share $\beta_{kk} = \frac{1}{1+(M_{\mathcal{C}}-1)}$ and gives a share $\beta_{jk} = \frac{\kappa}{1+(M_{\mathcal{C}}-1)}$ to each other members of the cartel, then the objective function of firm k is now

$$\tilde{\Pi}(q_k) = \Pi_k(q_k) + \kappa \cdot \sum_{j \in \mathcal{C} \setminus \{k\}} \Pi_j(q_k) \tag{2.18}$$

where the parameter $\kappa \in [0,1]$ controls the intensity of the collusion. The case where $\kappa = 1$ corresponds to the highest collusion intensity: cartel members fully internalize the impact of their pricing decisions on each other and maximize their joint profits. The case with no collusion corresponds to $\kappa = 0$. We therefore study the effect of moving for competition to collusion by taking the derivative with respect to κ. The demand side of the market is unchanged and firms face the same inverse demand functions as in the baseline case $\frac{\tilde{P}_k}{P} = \left(\frac{q_k}{y}\right)^{-\frac{1}{\rho}} \cdot \left(\frac{y}{c}\right)^{-\frac{1}{\eta}}$.

Proposition 8 (Markups under Collusion)**.** *The equilibrium price $\tilde{P}_k$ of a firm k that is part of a cartel is characterized by*

$$\tilde{P}_k = \tilde{\mu}_k \cdot \frac{W}{z_k}$$

where the own-elasticity depends on the combined market shares

$$\frac{1}{\tilde{\epsilon}_k} = \frac{1}{\rho} + \left(\frac{1}{\eta} - \frac{1}{\rho}\right) \left[s_k + \kappa \cdot \sum_{j \in \mathcal{C} \setminus \{k\}} s_j \right]$$

The result follows directly from the maximization problem. Colluding firms in the same sector internalize part of the effect of their decisions on the other cartel members' profits. Figure 2.2 illustrate the effect of collusion on the prices, markups, and market shares in the special case where $\kappa = 1$. In that case, firms in the cartel set uniform weights on all the cartel members' profits. As a result, they all choose the same markup. However, since they still have different productivities, their prices are different. The effect of collusion on the prices is clarified by the following proposition.

[10] Another view is to interpret price-fixing as a mutual agreement whereby cartel members share part of the *control* of their firms with other.

Proposition 9 (Sectoral Effects of Collusion)**.** *Under collusion, i) the prices and markups of the colluding firms increase; ii) some of the competitors who are not in the cartel increase their markups as well (Umbrella Pricing); iii) the total output of the colluding sector decreases; iv) firms lose market shares iff they increase their prices more than the increase in the price index.*

The first result follows from the fact that the colluding firms increase their markup, which raises their price. The first order effects of collusion on price is

$$d\ln P_k = \mu_k \left(\frac{1}{\eta} - \frac{1}{\rho}\right) \cdot \sum_{j \in \mathcal{C}\setminus\{k\}} s_j \cdot d\kappa$$

Because the overall price index is higher, equation (2.3) implies that consumers will substitute away from this sector. The magnitude of substitution is controlled by the between sector elasticity η , which is close to 1 in the baseline case. See Appendix C for the detailed proof.

Cartels also dampen the pass-through of productivity gains to prices. This is because firms that experience an increase in productivity gain market shares and are able to charge higher markups in return thereby reducing the pass-through of productivity shocks to prices. This is even more pronounced for cartel members, which can be seen from the cross-price elasticity matrix $\Omega[\kappa]$ defined as follows

$$\mathbf{\Omega}_{kj}[\kappa] = \begin{cases} \mathbf{\Omega}_{kj} & k \notin \mathcal{C} \\ \mathbf{\Omega}_{kj} + \kappa \sum_{i \in \mathcal{C}\setminus\{k\}} \left(\frac{\mu_k}{\mu_i}\right) \mathbf{\Omega}_{ij} & k \in \mathcal{C} \end{cases} \tag{2.19}$$

The cross-elasticity is unchanged for the firms not part of the cartel ($k \notin \mathcal{C}$) while there are extra terms for the firms in the cartels ($k \in \mathcal{C}$). These extra terms are the consequences of cartel members partially internalizing the price effects that they have on each other. The extra term is composed of the sum of the cross-price elasticities on the other cartel members, rescaled by their relative markups and by the collusion intensity.

Proposition 10 (Dampened Pass-through)**.** *The pass-through of productivity gains to prices is smaller in the presence of a cartel. In particular, we have*

$$0 < \|\mathbf{T}[\kappa]\|_\infty < \|\mathbf{T}\|_\infty < 1$$

where $\mathbf{T}[\kappa] \equiv [\mathbf{I} - \mathbf{\Omega}(\kappa])^{-1}$.

2.3.3 Aggregate Welfare

Because there is a continuum of sectors, in the current model, any increase of the price index in one single sector washes out in the aggregate. To build up an aggregate effect we consider deviations in a non-trivial measure of sectors, ϕ. We estimate ϕ by computing the weighted average of the sectors for which cases where found. We then simulate 1000 times the model where colluding firms are randomly drawn.

2.4 Data Description and Institutional Background

We assemble a new firm-level dataset on anti-competitive practices of French firms over the period 1994-2015. The information we collect relies on antitrust decisions taken by the French Competition Authority over the last 20 years. Cases spanning multiple countries are handled at the supra-national level by the Directorate-General for Competition of the European Commission. In this section, we describe important institutional details and the datasets we use.

2.4.1 Antitrust Decisions

Institutional Background. Antitrust regulation in France can be broken down into four periods, during which the competition regulator changed its name, structure, and mission. Born in 1953,[11] the French Technical Commission for Collusions and Dominant Positions' main goal was to fight against cartels and price fixing given their prevalence in post-war France. In 1963, its objectives were extended as the Commission also started investigating cases of dominant positions.[12] In practice, this Commission was directly notifying the Ministry for the Economy which would then decide whether to impose fines or not.

[11] "Décret n°53-704 du 9 août 1953".

[12] "Loi n°63-628 du 2 juillet 1963".

Following the 1973 oil crisis, Raymond Barre, an Economics Professor, was then Prime Minister in 1976 and advocated restricting even further price fixing arising from anti-competitive behaviors. In 1977, the Commission became the Competition Commission ("Commission de la Concurrence"). In addition to having to detect cartels and cases of dominant positions, the Commission was able to directly advise the French government on any competition-related matters but also on vertical and horizontal mergers and acquisitions.

The period 1986 to 2009 is important as it spans the beginning of our empirical analysis. Over this period, the Commission undergoes important transformations: its name is changed to the Competition Council ("Conseil de la Concurrence") and the 1986 Ordinance introduces several changes. Companies can directly refer cases to the Council. Moreover, the antitrust body becomes more independent, better protects concerned parties' rights and is now able to directly fine the firms found guilty of anti-competitive practices, though this does not apply to merger projects. The 2001 New Economic Regulation Law further introduces leniency and transaction programs to better detect and fight cartels.[13]

Finally, as of 2008, the Competition Council turns into the Competition Authority ("Autorité de la Concurrence" or ADLC, henceforth). The 2008 Law on the Modernization of Economy not only gives the right to the Authority to review merger and acquisitions independently from the Minister of Economy, but also to investigate potential anti-competitive cases on its own.

Anticompetitive Practices. As mentioned above, after investigating, the Competition Authority fines companies that are found guilty of engaging in any form of anti-competitive practice (abuse of dominant position, collusion or predatory pricing). Collusive behaviors might involve firms trading information on their prices and markups, imposing standard form contracts, enforcing barriers to entry, imposing exclusive or selective distribution agreements, market sharing, purposely stepping down from call for bids.

[13]A firm part of a cartel can go to the authority and report it. Under specific circumstances the firm will receive a more lenient fine that the other members of the cartels or not be fined at all. There are eleven cases involving the leniency program over the period 1994-2015.

The ADLC makes use of two tools in order to deter firm from taking part in illegal activities. The first one consists of fining these companies. The fines are set "according to the seriousness of the facts, the extent of the harm done to the economy, the individual situation of the company that has committed the infringement and of the group to which it belongs to, and whether it is an infringement that has been repeated or not".[14] The fines are capped as they cannot be higher than "10% of the global turnover of the group to which the company that is being fined belongs to". If the infringement is not committed by a company, the maximum amount of the fine is 3 million euros.[15] The second tool relies on issuing an injunction whereby the ADLC notifies the companies to change their behavior.

In practice, the information we extract to create our database comes from PDF files containing the description of the decisions made by the French Competition Authority. These files are freely available on the ADLC website. We make use of an automatic textual analysis to retrieve information on the identity of the firms fined by the antitrust body.[16] Crucially, our database contains the name of the firms that are fined which signals that these companies behaved illegally and are anti-competitive. We also have information on the amount of the fine in thousands of euros, the year the verdict is returned and the starting year of the investigation. We then use the companies' names to back out their national identification code ("SIREN" code) given by the French National Institute of Statistics and Economic Studies (INSEE). This allows us to match our database with other firm-level production datasets. We describe in the Data Appendix how we assemble our database, the variables it contains and the features we will add in future versions of our paper.

Timing Assumptions. The information we do not have access to in the current version of the paper is the duration of cartels and how long did each firm engage in anti-competitive behavior. Getting data on the exact starting date of the cartel is complicated and the

[14]French Commercial Code, L.420-1 or L.420-2.

[15]French Commercial Code, L.464-2.

[16]We do not use information on firms notified by an injunction. Often, these firms are fined later on by the ADLC and thus appear in our database.

reported date is likely to be inaccurate as companies have an incentive to misreport the birth date of the cartel to pay a lower fine. We thus make one assumption on the duration of cartels in order to estimate the impact of anti-competitive behaviors on firm-level outcomes. We assume that firms fined by the ADLC were anti-competitive for a period of five years before being fined by the ADLC. The duration of five years corresponds to the average duration of discovered French cartels observed over the period 2003-2015 by **monnier**. This also matches the average duration of cartels summarized in Levenstein and Suslow 2006 for a wide range of studies.[17]

One can argue that firms stop colluding before the ADLC verdict is returned. As a robustness check, we change the timing of collusion and assume that firms collude for a period of five years before the ADLC is notified of the illegal behaviors. These assumptions might seem extreme to the extent that they entail that firms stop colluding right when they are being fined or investigated. This might not be the case if the benefit of joining and staying in a cartel is higher than the cost of getting caught and sanctioned. Moreover, Stigler 1964 argues that a firm in a cartel has an incentive to deviate and price below its competitors to increase its market share. This leads him to conclude that cartels are by nature unstable. Empirically, this does not seem to be the case **monnier**; Levenstein and Suslow 2006 and our assumptions are not ill-advised: as **monnier** show, around 63% of French cartels die after the ADLC starts investigating or when a company or client files an official complaint and seizes the Competition Authority.

One caveat is that we are making the implicit assumption that the date at which the ADLC was notified corresponds to the start of the investigation and that the companies are fully aware of it. In practice, it takes more time for the ADLC to launch an investigation and notify the potential colluders. The duration of cartels will be wrongly assigned if the year when the ADLC was notified of the case is different from that when the companies were

[17]It is possible to argue that these dates are not a correct approximation of the true duration of cartels as these studies are based on discovered cartels that might not be representative of all cartels. Harrington Jr and Wei 2017 study in a theoretical framework the magnitude of the bias associated to this issue. They find the bias to be modest which lead them to conclude that using the duration of discovered cartels as a proxy for the duration of all cartels, discovered and undiscovered, is not a bad approximation.

notified of the start of the investigation. Finally, to the extent that there is a lot of variation in cartels duration, with some lasting for a few days and some lasting for several years, we further assume in a robustness check that firms engage in anti-competitive practices for two years, corresponding to the first-quartile of cartels duration according to **monnier**. Our results are quantitatively and qualitatively unaltered. In a future version of our paper, we will add more precise information on the duration of cartels from the decision files.

2.4.2 Firm-level datasets

We match our database on anti-competitive firms with firm-level data for France. Matching the firms is made possible by the fact that French firms are assigned a unique identifier ("SIREN" code). The datasets that we use contain the universe of French firms over the period 1994-2015.[18] These datasets contain the balance sheets and income statements of all French firms. We keep both large and smaller firms which corresponds to two different tax regimes, the Regime of Normal Real Profits (BRN) and the Simplified Regime for the Self-Employed (RSI), respectively. BRN contains firms with annual sales above 763K euros (230K euros for services) whereas smaller firms included in RSI sell at least 76.3K euros (but less than 763K euros) a year and more than 27K euros for services. However, BRN is the most relevant data source given that in 2003, BRN firms' sales share in total sales was 94.3% and is constant over time. This data has been used in previous studies, for instance in Di Giovanni, Levchenko, and Mejean 2014 and we refer to their paper for more details. Importantly, these exhaustive databases allow us to build a firm's labor share, market share and other variables we directly use in our empirical framework. More information on the variables we use and build is provided in the Appendix.

Finally, we complement FICUS-FARE by making use of the LIFI survey database. This database has been used previously in Giovanni, Levchenko, and Mejean 2018, for instance. This survey provides us with the ownership status of firms. More specifically, we use it to know whether firms are owned or not, the identity of the owner and its origin (foreign or

[18]For the period 1994-2007 and 2008-2015, we rely on FICUS and FARE, respectively.

domestic). This survey is important to us as it allows us to know whether anti-competitive firms are independent and make their own decisions or not. Because the decision to assign which firm is independent is somewhat arbitrary, we rely on different thresholds and samples and show that our results are robust to using these various methods. More details on how we create our final sample can be found in the Appendix.

2.4.3 Definitions

In FICUS-FARE, each firm is assigned a 5-digit principal activity code ("Code APE") by the INSEE and whose aim is to pin down in which industry the firm mostly operates. Because the precise breakdown of sales across products is not available for the French data, the relevant market for a firm is its 5-digit industry code. Therefore, throughout the paper, we will denote a firm's market share by its market share in the relevant 5-digit industry code.

Our definition of sector follows the NAF Rev. 2 classification. There is a many to one matching between the 5-digit APE codes and the 2-digit NAF Rev. 2 sectors. Every firm is thus assigned to a 2-digit sector.

2.4.4 Concentration and Anti-Competitive Firms

There has been an increase in US concentration documented in several recent papers Autor et al. 2017; Gutiérrez and Philippon 2017. Concentration is usually measured by studying the evolution of the Herfindahl-Hirschman Index (HHI) or concentration ratios (CR). In Figure 2.3, we plot the evolution of the CR4 and CR20 ratios over 1994-2015. The concentration ratio is computed for each 5-digit industry and we then take the mean of all the industry CR at a given point in time. The pattern is clear: concentration has also increased in France.

Table 2.1 shows some descriptive statistics on anti-competitive cases. The third and fourth column display the share of each sector's sales and gross value-added while the fifth column shows the average number of anti-competitive firms in each sector over the period 1994-2015. There are two sectors in which no firm was convicted, namely the agricultural and the education sectors. Cartelization and more generally anti-competitive practices are

prevalent in France and most cartels involve firms operating in the construction, wholesale and retail and transportation sectors. These three sectors account for almost 50% of total sales and 36% of total value-added in France.

In Table 2.2, we investigate the characteristics of anti-competitive firms versus firms that have not been officially sentenced. The Colluding firms have a much higher market share on average: 4% versus 0.007% for non-colluders. The average cumulated market share of these colluders is equal to 11%: on average, these firms represent a non-trivial fraction of a market's total sales. Colluders also sell more, spend more on intermediate goods, have more employees, use capital more intensively and are more productive, as measured by revenue TFP and labor productivity.[19] These summary statistics are most-likely the result of self-selection into colluding, whereby more productive and bigger firms are more likely to find it profitable to join a cartel Bos and Harrington 2010. However, our empirical framework aims to shed light on the treatment effect of anti-competitive behaviors and cartelization.

2.5 Empirical Framework

In this section, we describe the different econometric specifications we take to the data and potential threats to identification that we will address in a future version of our paper.

2.5.1 Anti-Competitive Firms and Firm-Level Outcomes

In the present version of our paper, we do not aim to directly test our theoretical predictions. Instead, we provide some evidence on the relationship between the existence of cartels and firm-level outcomes such as market shares and employment.

The main specification we estimate is:

$$y_{kt} = \beta_1 \text{AntiComp}_{it} + \beta_2 \text{AntiComp}_{it} \times \text{Colluder}_{kt} + \mathbf{Z}'_{kt}\theta + \psi_{It} + \nu_i + \varepsilon_{kt} \tag{2.20}$$

The variable y_{kt} contains the market share, (log) employment or (log) wages of a firm k in t. AntiComp_{it} captures the existence of colluding firms in a 5-digit industry i. This is a

[19] We detail the estimation procedure used to back out revenue TFP in the Appendix.

dummy equal to one if at least one firm was fined by the ADLC in an industry i. We choose the threshold value of one because firms may abuse their dominant position and this type of infringement only requires the existence of one company. Colluder_{kt} is an indicator variable that takes the value 1 if firm k was fined by the ADLC. It is important to note that at this stage, we will therefore not capture whether firms belonging to the *same* cartel gain market shares and the effect on non-colluders within the same industry. $\mathbf{Z}'_{kt}$ is a set of controls, ψ_{It} are 2-digit sector-year fixed effects that control for sectoral shocks that might affect all firms in the same manner, while ν_i are 5-digit industry fixed effects. The intuition for this specification is the following: the existence of colluders can affect firms' market share but the effect might be differentiated depending on whether the firm behaves anti-competitively or not. This differentiated impact is captured by the interaction term.

The advantage of this specification is that we control for the fact that some industries might be more concentrated and as a result more prone to cartelization. For instance, specific industries in which barriers to entry are always high and limit the number of competitors might make it easier for firms to coordinate and gain market shares by colluding Bain 1956. However, these 5-digit industry dummies will not capture the intensity of the anti-competitive behaviors as industries with a high number of anti-competitive companies will be counted in the same way as industries with a single firm, as big or small that firm might be. We will therefore also include the total number of anti-competitive firms within a 5-digit industry to control for the fact that some industries are more severely impacted by colluding cases.

Upstream/ Downstream Linkages We also seek to identify how cartels and anti-competitive practices more generally propagate in the economy and affect downstream and upstream sectors. To do so, we will first make use of French Input-Output (I-O) Tables to build on Antràs et al. 2012.[20] Their methodology allows us to build an upstreamness index that we can then interact with the dummy variable for whether the firm is anti-competitive

[20]As opposed to the US I-O Tables, the French I-O Tables are fairly more aggregated: depending on the period considered, they contain 35 or 56 sectors.

or not in (2.20).[21]

We will then estimate the direct effect of cartels on a firm's outcome operating in the same sector as that cartel but taking into account the fact that some sectors upstream and downstream might contain cartels. We then test if the shocks in other sectors have an effect.

2.5.2 Identification

One might worry that our analysis suffers from sample selection bias because anti-competitive firms in our sample are *discovered* firms. The issue is that there might be a myriad of other colluding companies that go unnoticed and that might be/behave fundamentally differently from discovered firms, affecting our results. Undiscovered cartel members would be classified as competitive but their characteristics and behavior might lead them to extract more market power. While this is a possibility that is impossible to rule out, we argue that, if anything, this would lead us to *underestimate* the effect of competition, or lack thereof, on our firm-level outcomes such as employment and market shares: firms assumed to be competitive that in fact collude are able to extract market power via secret agreements.

Moreover, the likelihood that firms self-select into colluding cannot be rejected: larger, more productive firms might have a strong incentive to form or join a cartel Bos and Harrington 2010. If this is the case, our empirical results will not reflect any treatment effect but rather a selection effect. We plan to pursue our line of research by using propensity score matching (PSM) to create a control group similar to colluding firms based on some observable characteristics such as revenue TFP, capital intensity, employment, firm level wages etc. We can then estimate an average treatment effect of colluding on firm-level outcomes by comparing our treated group to the newly created control group. Although this methodology has been used recently by Blonigen and Pierce 2016 to measure the effect of mergers and acquisitions on product market power and revenue TFP, we are not aware of other studies trying to use this method to assess the effect of anti-competitive practices on

[21]The lower the upstreamness index, the closer that sector is to final demand. For instance, we expect the motor vehicles, trailers and semi-trailers sector to have a low index but the coke and refined petroleum products sector to have a high upstreamness index.

firm-level outcomes.

2.6 Results

In this section, we discuss the effect of colluding behaviors on firm-level outcomes. As mentioned before, the estimates are not causal.

Main Results. Table 2.3 displays our baseline results. The existence of anti-competitive firms within an industry leads to a drop in market share of 0.06 percentage points for firms that do not collude. Colluding firms benefit from it and their market share increases by 3.8% percentage points when there is at least one colluder in that industry.Although the effect drops to an increase of 2.8% percentage points when we control for firm size and revenue TFP, the effect remains highly significant. At first sight, anti-competitive firms therefore seem able to gain market shares. As discussed, the effect might be driven in part by the fact that some industries are more concentrated due to barriers to entry, for example. We therefore add 5-digit industry dummies in column (3). While the effect on AntiComp_{it} is now one order of magnitude lower, firms that behave illegally still largely benefit from it and this still holds when we further control for revenue TFP in column (4). In the last column of the table, we also control for the intensity of within-industry collusions by adding the total number of colluding firms in an industry and interact it with the colluder dummy. The interaction term is negative and highly significant. As expected, the higher the number of anti-competitive firms in an industry, the lower the market share of colluding firms everything else equal. However, the total effect of acting illegally on its market share remains positive and significant.

We also test whether the effect is driven by relatively more productive firms. To do so, we assign firms in four different quartiles, depending on their place in the productivity distribution. Apart from the bottom 25%, Table 2.4 shows that relative more productive firms that collude are able to extract more market power in the presence of other colluding firms in their industry. While the interaction term is unusually high for the first quartile, the

point estimate becomes increasingly high as we move from the second to the last quartile. Though not displayed, we also tested how bunching the firms in quartiles depending on their (log) capital intensity changes our results and we found similar qualitative evidence. Finally, Table 2.5 shows how colluding affects a firm's employment and wages. The effects are the same qualitatively for employment. Quantitatively, a sectoral collusion decreases a firm's employment by 0.0013 percent but it increases a colluder's employment by 0.02 percent (first column). This might be because colluders are large firms that also dominate the labor market. The estimates are robust to controlling for the number of colluders in each industry. The effect on wages is not fully clear as shown in the last two columns. The point estimates on the industry dummy switch sign as we control for the number of colluders and are not significant anymore.

Robustness. We provide evidence that our results are robust to our timing assumption and to changing the identity of the colluder. In Table 2.6, we change the timing assumption as described in Section 4.1. In the first column, we keep the average duration of five years but we assume that firms stop colluding when the ADLC is notified of the case. As an example, this amounts to assuming that a cartel member whose cartel was reported to the ADLC in 2007, was colluding from 2003 to 2007. The results are qualitatively similar and very close in terms of point estimates. Similarly, in the last two columns, we assume that the duration of collusion is equal to two years, corresponding to the lowest quartile found by **monnier**. Once again, although the estimates slightly change in terms of magnitude, they remain significant and very close to our baseline results. In Table 2.7, we assume that the fined firm is not the colluding firm anymore. Instead, we make use of the LIFI database and assign the colluding dummy to the owner of that firm. The results remain very close to those found in Table 2.3 with a coefficient on AntiComp significant at the 10.2% level in the fourth column and on the last interaction in Column (5) significant at the 10.3% level. Therefore, our results do not seem to be driven by our timing assumption or by the identity of the colluder.

2.7 Conclusion

We develop an oligopolistic framework with heterogeneous firms and endogenous markups where firms can form cartels. Collusive practices operate like patterns of cross-ownership and distort the firms' incentive. This framework allows us to derive analytically the impact of cartelization. We simulate and test the predictions of our model using a new firm-level database by extracting information from the French Competition Authority's decisions. In this preliminary version, several important parts have been omitted that will be added in future versions. We plan to complete the welfare analysis part of our model, update and complete our firm-level database on anti-competitive conduct and improve our empirical framework to causally estimate the effect of market structure distortions on firm-level outcomes.

Supplementary Figures

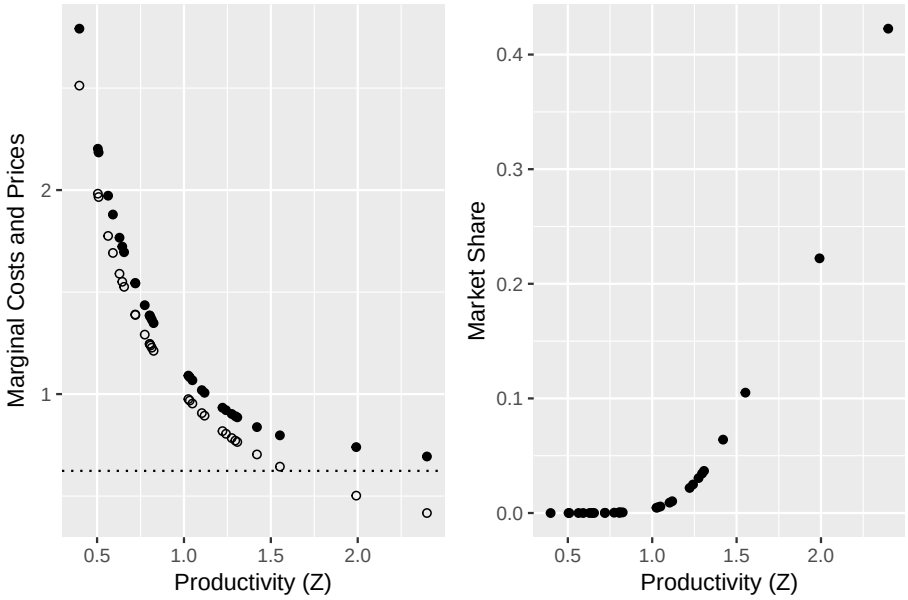


Figure 2.1: Oligopolistic Industry Equilibrium

Firms are ranked in increasing productivity order on the horizontal axis. The left panel shows the marginal costs (circles) decrease with productivity. The prices (black dots) also decrease with productivity but lie above the marginal costs. The horizontal dotted line represent the sector's price index. The right panel traces the profile of the markup function.

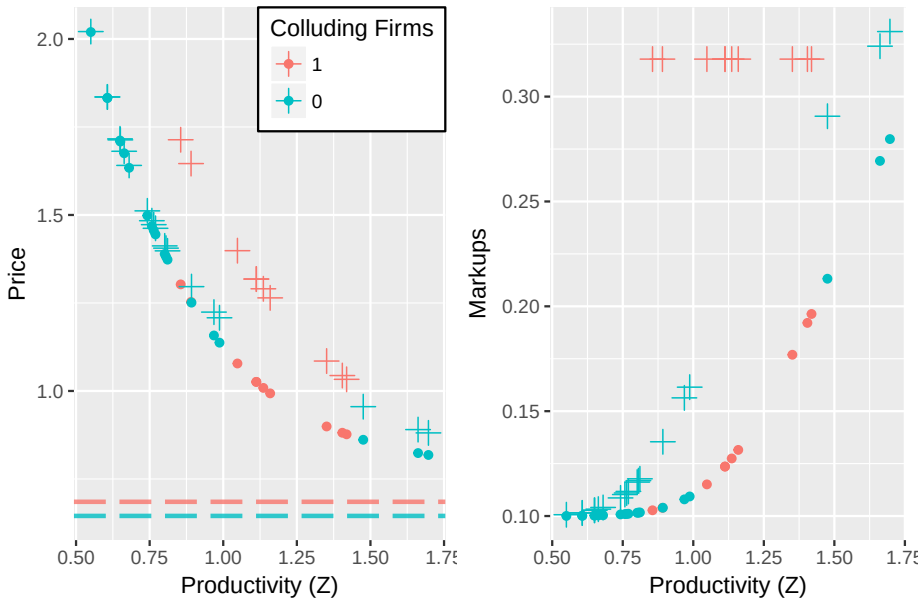


Figure 2.2: Oligopolistic Industry Equilibrium with collusion

Firms in the cartel raise their prices (red crosses) with respect to the baseline (dots). Firms who are not part of the cartel also raise their prices (green crosses). This is the umbrella pricing effect. The price index in the industry goes up, decreasing consumer's welfare. In this example we set $\kappa = 1$, therefore cartel members all choose the same markup.

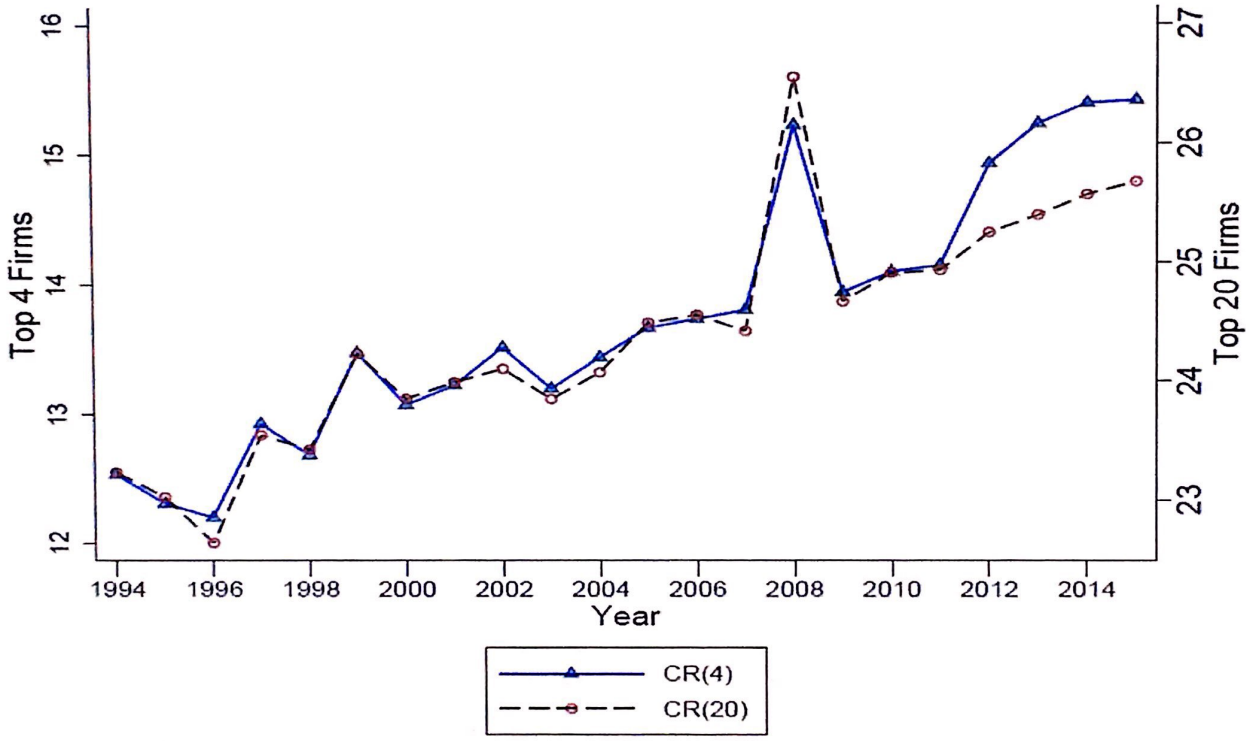


Figure 2.3: Concentration Ratios

Data Source: FICUS-FARE.

Supplementary Tables

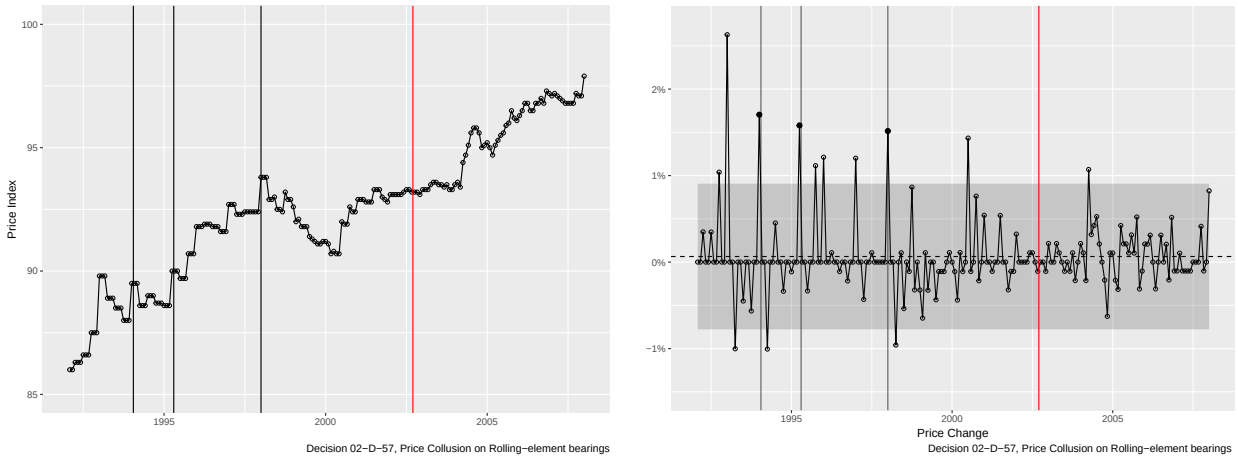


Figure 2.4: Example of Price Fixing in the Ball bearings Industry

Concerted price increases (black vertical lines) were revealed by the investigation. The red line represents the time when the Competition Authority gave its sentencing decision. The dark grey area represents the +/- 2 standard deviations of the mean monthly price change.

2.A Data Appendix

2.A.1 Firm-level Database on Anti-competitive Conduct

In order to extract information on the identity of the firms fined by the ADLC we proceed as follows. First, we scrape the website of the ADLC to recover all the decision files over the period 1994-2015. These PDF documents contain information on the situation of the market impacted by anti-competitive behaviors, the notification date of the case to the ADLC, the names of the firms fined for anti-competitive behaviors, the types of infraction they committed, their sales and sometimes an estimate of the duration or date of the infraction. Some of these files contain information on when the firms were notified by the ADLC that an investigation is going to be launched. Extracting and getting data on the identity of these anti-competitive companies is straightforward to the extent that the layout is relatively similar across decision files. A salient and important example is that of the companies' name which always appear at the end of the PDF right after the word *Décide* ("Decides").

Second, for the moment, we use Python's textual analysis tools to back out the name of these companies, their sales, the date when the ADLC was first notified of the infraction and

Table 2.1: Collusions by Sector

NAF	Sector	Sales (Share)	VA (Share)	# Collusions
01-03	Agriculture, forestry and fishing	0.0010	0.0014	0
05-09	Mining and quarrying	0.0027	0.0038	0.82
10-12	Food products, beverages and tobacco	0.0558	0.0491	10.55
13-15	Textiles, apparel, leather and related prod.	0.0108	0.0114	0.55
16-18	Wood and paper prod., and printing	0.0164	0.0178	3.27
19	Coke, and refined petroleum prod.	0.0187	0.0171	1
20-21	Chemicals and chemical products	0.0402	0.0381	5.27
22-23	Rubber and plastics prod., and other non-metallic mineral prod.	0.023	0.027	6.41
24-25	Basic metals and fabricated metal prod., except machinery and equipment	0.0295	0.0326	6.68
26-27	Electrical and optical equipment	0.0247	0.0265	2.91
28	Machinery and equipment n.e.c.	0.0163	0.0171	1.59
29-30	Transport equipment	0.052	0.037	1.5
31-33	Other manufacturing; repair and installation of machinery and equipment	0.016	0.021	1.77
35-39	Electricity, gas, and water supply	0.034	0.0437	4
41-43	Construction	0.064	0.081	48.14
45-47	Wholesale and retail trade, repair of motor vehicles and motorcycles	0.3671	0.1933	48.55
49-53	Transportation and storage	0.0538	0.083	30
55-56	Accommodation and food service activities	0.0205	0.0333	0.95
58-60	Publishing, audiovisual and broadcasting activities	0.0152	0.024	2.68
61	Telecommunications	0.0169	0.0311	1.77
62-63	IT and other information services	0.0147	0.0266	0.86
68	Real estate activities	0.0132	0.0262	0.86
69-82	Professional, scientific and administrative activities	0.0707	0.1206	15.41
85	Education	0.002	0.0037	0
86-88	Health and social work	0.0081	0.0173	5.91
90-93	Arts, entertainment and recreation	0.0068	0.0059	0.32
94-96	Other service activities	0.0056	0.010	2.77

Notes: The Sales (Share) column represents sector-level sales in total sales over the period 1994-2015. The VA (Share) column represents sector-level value-added in total value-added over the period 1994-2015. The values displayed for the number of collusions are averages over the period 1994-2015. Colluding firms are assumed to be colluding for a period of 5 years before the decision of the ADLC as explained in Section 4.

Table 2.2: Summary Statistics

	Full Sample		Colluding Firms		Non Colluding Firms	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Market Share	0.000725	0.010	0.0397	0.12	0.000716	0.0099
Cumulated Market Share	X	X	0.1145	0.163	X	X
Sales	2857.859	94074.28	440436.8	2585997	2760.201	85539.76
Value-added	815.74	33044	198115.3	1416515	771.70	25214
Revenue TFP	1.842	0.43	1.96	0.37	1.84	0.43
Labor Productivity	3.77	0.72	4.15	0.78	3.77	0.72
Labor	14	454	2331	15985	14	385
Capital/Labor ratio	0.302	18.14	1.458	23.67	0.302	18.134
Intermediates	2064.66	72512	252997	1451346	2008.66	69102
# Obs.	20,167,666		4,500		20,163,166	
# Firms	2,884,325		1005		2,884,305	

Notes: The values displayed are for the period 1994-2015. Sales and value-added are in thousands of euros. Revenue TFP is obtained by estimating a production function by sector as described in Appendix B. Labor productivity is real value-added (deflated by 2-digit price indices) divided by the number of workers. Labor is the number of workers. The capital-labor ratio is expressed in real terms where capital has been deflated. Intermediates is the value of expenditures on intermediate goods in thousands of euros. The number of colluding and non-colluding firms does not add up to the total number of firms in the full sample because firms do not always collude and therefore switch from one status to the other.

the corresponding amount of the fine for each firm. This step requires some manual cleaning as some companies, numbers and cases are misreported. We therefore go through all the files to complement the information extracted from the textual analysis and double check that

Table 2.3: Baseline Results

Dependent Variable	Market Share$_{kt}$				
	(1)	(2)	(3)	(4)	(5)
AntiComp$_{it}$	-0.06***	-0.068***	-0.006**	-0.00544*	-0.0052**
	(0.02)	(0.02)	(0.0029)	(0.0027)	(0.0024)
AntiComp$_{it}$ × Colluder$_{kt}$	3.9***	2.9***	3.7***	3.24631***	5.679***
	(0.94)	(0.89)	(0.91)	(0.95)	(1.406)
Revenue TFP$_{kt}$		0.055		0.15**	
		(0.043)		(0.062)	
# Colluders$_{it}$					-0.0005
					(0.0007)
# Colluders$_{it}$× Colluder$_{kt}$					-0.262***
					(0.09)
2-digit Sector × Year FE	✓	✓	✓	✓	✓
5-digit Industry FE			✓	✓	✓
Firm Size Controls		✓			
# Obs.	20,167,666	15,040,842	20,167,662	15,040,836	20,167,662

Notes: Standard errors clustered at the 2-digit sector level. *** p<0.01, ** p<0.05, * p<0.1. Firm size controls include the quantity of workers, capital and materials (in log).

our newly created dataset is not missing anything that would appear in the original PDF files but that we would miss via the textual analysis exercise. At this stage, the dataset is informative about the identity (name) of the firms that were fined by the French Antitrust Authority, their sales, the case number of the decision, the amount of the fine for each firm and the notification date of the case to the ADLC.

Third, we make use of Orbis and Python to recover information on the identification

Table 2.4: Results by Productivity Quartile

Dependent Variable	Market Share$_{kt}$			
Sample	$z_{kt} \in \text{Q}_1$	$z_{kt} \in \text{Q}_2$	$z_{kt} \in \text{Q}_3$	$z_{kt} \in \text{Q}_4$
AntiComp$_{it}$	-0.00595	-0.00213	-0.00432***	-0.0139**
	(0.00588)	(0.0.00214)	(0.00151)	(0.00626)
AntiComp$_{it}$ × Colluder$_{kt}$	6.058*	1.484*	1.978*	3.92***
	(3.07)	(0.752)	(0.964)	(1.207)
Revenue TFP$_{kt}$	0.078	0.119**	0.174***	0.184
	(0.068)	(0.0475)	(0.0602)	(0.168)
2-digit Sector × Year FE	✓	✓	✓	✓
5-digit Industry FE	✓	✓	✓	✓
# Obs.	3,760,204	3,760,194	3,760,183	3,760,181

Notes: Standard errors clustered at the 2-digit sector level. *** p<0.01, ** p<0.05, * p<0.1.

number of the firms which will then allow us to match our database to the balance-sheets data. To do so, we upload our temporary database into the Batch Search engine of Orbis to look for the SIREN number of each firm given its name. We complement this information with a Python script that allows us to obtain the SIREN number of firms based on a Bing search of that firm's name.[22] Although these methods are imperfect, they facilitate the matching with FICUS-FARE.

Finally, before matching our database with FICUS-FARE, we manually verify that the

[22]We thank Arthur Guillouzouic Le Corff for sharing his code.

Table 2.5: Effects on Employment and Wages

Dependent Variable	$\ln$ Employment$_{kt}$		$\ln$ Wages$_{kt}$	
	(1)	(2)	(3)	(4)
AntiComp$_{it}$	-0.013*	-0.015**	0.003*	-0.0001
	(0.0067)	(0.007)	(0.0019)	(0.0028)
AntiComp$_{it}$ × Colluder$_{kt}$	2.375***	2.931***	0.091*	0.307***
	(0.25)	(0.195)	(0.052)	(0.025)
Revenue TFP$_{kt}$	0.254	1.18		
	(0.38)	(0.89)		
# Colluders$_{it}$		0.00008		0.0002
		(0.0013)		(0.00015)
# Colluders$_{it}$× Colluder$_{kt}$		-0.055***		-0.0067***
		(0.0068)		(0.0015)
2-digit Sector × Year FE	✓	✓	✓	✓
5-digit Industry FE	✓	✓	✓	✓
# Obs.	15,040,836	15,037,090	20,167,662	20,145,535

Notes: Standard errors clustered at the 2-digit sector level. *** p<0.01, ** p<0.05, * p<0.1.

SIREN numbers obtained from Orbis and from our scraping procedure are correct. We do so by making sure that the sales (in euros) of the firm in our database correspond to those reported in FICUS-FARE. For the firms that were not matched by any means in our third step, we manually search for them in FICUS-FARE using the information on their sales and

add their SIREN number directly in our database. For now, we create the full database over the period 1994-2015 by making the assumption that the average duration over which firms behaved anti-competitively is five years. This corresponds to the average amount found over the period 2003-2015 in France by **monnier** and is consistent with other cartel cases in different settings Levenstein and Suslow 2006, as explained in detail in Section 4.1 of the paper.

In future versions of our paper, we will update this database by adding more detailed information on the duration of discovered cartels for the cases for which this information is available, add firms that were warned by the ADLC but not fined, add the total number of firms discovered in each cartel, add information on whether the firms colluded in prices or quantities.

2.A.2 List of Variables

We describe below the different variables used in our empirical framework. Note that our main sample consists of observations with strictly positive values for gross value-added, total and domestic sales, number of employees, labor compensation, expenditures on materials and capital.

- **Anti-Competitive Industry:** For each 5-digit industry in a given year, we count the total number of colluding firms and create a dummy variable equal to one if there is at least one firm in that industry. We choose a value equal to one because firms can abuse their dominant position. *Source: Moreau-Panon database*

- **APE Code:** 5-digit industry code. Before 2008, APE codes are available in a 4-digit format corresponding to the NAF Rev. 1 classification. We convert all these NAF Rev. 1 codes into Naf Rev. 2 codes using a correspondence table available on the INSEE website. Our matching procedure is such that each of the 712 NAF Rev. 1 APE code is assigned to one NAF Rev. 2 APE code. Our code is available upon request. *Source: FICUS-FARE and authors' calculation*

- **Capital:** Net book value of capital. We cannot build a capital measure using the perpetual inventory method as there is a break between FICUS and FARE and no data on investments is reported in 2008. We further deflate capital expenditures by sector-level price indices from EUKLEMS Jäger 2017. *Source: FICUS-FARE and authors' calculation*

- **Colluder:** Dummy variable that takes the value one if the firm engaged in anti-competitive practices in a given year. *Source: Moreau-Panon database*

- **Employment:** Total number of employees working in each firm. *Source: FICUS-FARE*

- **Export Sales:** Export sales reported by the firm in thousands of euros. This variable is available in the fiscal files and is highly correlated (correlation coefficient above 0.9) with total export sales computed from the customs data. Firms are classified as exporters if they sell a positive amount abroad according to the customs. *Source: Customs data and FICUS-FARE*

- **Gross Value-Added:** This variable is directly available in FICUS-FARE and follows the accounting definition according to which it is equal to total sales minus input expenses taking into account changes in inventories. *Source: FICUS-FARE*

- **Labor Compensation:** This variable is the sum of two components separately available in the fiscal files: salaries and social benefits that are paid by the employer and that benefit the worker in the form of retirement funds, social security funds etc. *Source: FICUS-FARE*

- **Labor Share:** Consistent with Karabarbounis and Neiman 2014, Elsby, Hobijn, and Şahin 2013, we construct our firm-level labor share variable as follows. In accounting, gross value-added is equal to the sum of gross operating surplus, labor compensation (as defined above) and taxes net of subsidies. We therefore do not allocate taxes net of subsidies and build the labor share as the ratio of labor compensation to gross value-

added. Observations with values outside the $(0, 1)$ interval are discarded whenever our sample involves labor shares. *Source: FICUS-FARE and authors' calculation*

- **Market Shares:** A firm's market share is defined at the 5-digit level. We compute market shares by dividing a firm's total sales by the total amount sold by all the firms operating in the same market at a point in time. *Source: FICUS-FARE and authors' calculation*

- **Materials:** Materials are defined as the sum of expenditures on raw materials, final goods and other categories. We further deflate this expenditure variable by 2-digit sector intermediate goods price indices from EUKLEMS. *Source: FICUS-FARE and authors' calculation*

- **NAF Code:** 2-digit sector code according to the NACE Rev. 2 classification. Some sectors are pooled together, depending to the availability of sector-price deflators. The list of sectors is displayed in Table **??**. *Source: FICUS-FARE*

- **Ownership Status:** [TBC] Describe method and deflators *Source: LIFI*

- **Total Sales:** Total sales (domestic sales plus export sales) reported by the firm in thousands of euros. *Source: FICUS-FARE*

- **Wages:** Firm-level wages are obtained by dividing labor compensation by employment. *Source: FICUS-FARE and authors' calculation*

2.B Estimation Appendix

Our estimation method relies on the seminal papers of Olley and Pakes 1992, Levinsohn and Petrin 2003 and Ackerberg, Caves, and Frazer 2015. The idea is that output is produced by using labor, capital, materials and productivity. Total factor productivity (TFP) is a residual because it is not observed and most importantly, its absence in standard production function estimation leads to biased estimates of labor, capital and materials. This is due to the fact that these inputs are chosen depending on the productivity realizations that the firm

observes Marschak and Andrews 1944. The way we control for it and back out productivity is by assuming that the demand for materials is a function of capital, labor and productivity as in Ackerberg, Caves, and Frazer 2015.

Formally, we assume a Cobb-Douglas production function in log-form where output y is being produced by labor l, capital k, materials m and depends on productivity ω which is Hicks neutral. To ease the notation, we index firms by i instead of k which denotes capital in the following application.

$$y_{it} = \alpha_l l_{it} + \alpha_k k_{it} + \alpha_m m_{it} + \omega_{it} \tag{2.21}$$

The Hicks neutral term ω_{it} is a function of a predictable term z_{it} that the firm has access to but is unobserved to the econometrician and a noise ξ_{it}. For simplicity, we assume that ω_{it} is the sum of these two components:

$$y_{it} = \alpha_l l_{it} + \alpha_k k_{it} + \alpha_m m_{it} + z_{it} + \xi_{it}$$

We follow Ackerberg, Caves, and Frazer 2015 and assume that materials is an invertible function of labor, capital and the only unobserved term z_{it}:

$$m_{it} = \Phi_t\big(l_{it}, k_{it}, z_{it}\big)$$

This implies that we can invert the demand for materials to control for productivity as a function of observables:

$$z_{it} = \Psi_t\big(l_{it}, k_{it}, m_{it}\big)$$

where $\Psi_t(.) := \Phi_t^{-1}(.)$. The resulting equation is:

$$y_{it} = \alpha_l l_{it} + \alpha_k k_{it} + \alpha_m m_{it} + \Psi_t\big(l_{it}, k_{it}, m_{it}\big) + \xi_{it}$$

which we rewrite as:

$$y_{it} = f_t\big(l_{it}, k_{it}, m_{it}\big) + \xi_{it} \tag{2.22}$$

The estimation method consists of two steps. In the first step we non parametrically estimate equation (2.22). In practice, we approximate $f_t(.)$ by a third order polynomial in its

arguments as well as interactions of all the terms. This gives us predicted output $\hat{f}_t(.)$. We then use the fact that:

$$z_{it} = \hat{f}_t\big(l_{it}, k_{it}, m_{it}\big) - \alpha_l l_{it} - \alpha_k k_{it} - \alpha_m m_{it} \tag{2.23}$$

We can now specify the law of motion of productivity which we assume follows a first-order Markov process:

$$z_{it} = h_t\big(z_{it-1}\big) + \vartheta_{it}$$

In practice, we estimate:

$$\hat{z}_{it}\big(\alpha_l, \alpha_k, \alpha_m\big) = \sum_{j=1}^{3} \beta_j \hat{z}_{it-1}^{j}\big(\alpha_l, \alpha_k, \alpha_m\big) + \vartheta_{it} \tag{2.24}$$

where we have made clear that productivity is derived from the estimation of (2.23) and a guess on $\big(\alpha_l, \alpha_k, \alpha_m\big)$. Estimating (2.24) gives us an estimate of $\vartheta_{it}\big(\alpha_l, \alpha_k, \alpha_m\big)$ which is the innovation term to productivity.

The second stage of the estimation procedure consists of using moment conditions and estimating the system by GMM:

$$\mathbb{E}\left(\hat{\vartheta}_{it}\big(\alpha_l, \alpha_k, \alpha_m\big)\begin{pmatrix} l_{it} \\ k_{it} \\ m_{it-1} \end{pmatrix}\right) = 0 \tag{2.25}$$

These moment conditions are standard in the empirical IO literature. Labor, capital are assumed to be dynamic inputs so that the innovation term is uncorrelated with their value at time t. Materials are assumed to be flexible so that their demand might vary with the innovation shock in t and we must use their lagged value instead. The parameters of interest solve the moment conditions in (2.25).

Once we have recovered the output elasticities, we can define productivity as the Solow residual:

$$\hat{z}_{it} = y_{it} - \hat{\alpha}_l l_{it} - \hat{\alpha}_k k_{it} - \hat{\alpha}_m m_{it} \tag{2.26}$$

2.C Mathematical Appendix

2.C.1 Properties of the Industry Equilibrium

This section formulates the industry equilibrium as a nested fixed point in the space of prices and derives the main results of the paper.

Lemma 1 (Nested Fixed Point)**.** *The vector of equilibrium prices,* $\mathbf{P} = (P_k)_{k=1,...,K}$ *is the unique solution to the following nested fixed point problem*

$$\begin{cases} P_k = \Phi\left(P_k; z_k, P\right) \ \ \forall k = 1, ..., K \\ P = \Psi\left(\mathbf{P}\right). \end{cases}$$

Proof: for ease of notation, we drop the sector subscript i. Rearranging the firms' first order conditions on prices, we have

$$\begin{aligned} P_k &= \frac{1}{1-\frac{1}{\epsilon_k}} \frac{W}{z_k} \\ &= \left[1 - \frac{1}{\epsilon_k}\right]^{-1} \frac{W}{z_k} \\ &= \left[1 - \frac{1}{\rho}\left(1 - s_k\right) - \frac{1}{\eta} s_k\right]^{-1} \frac{W}{z_k} \\ &= \left[1 - \frac{1}{\rho} - \left(\frac{1}{\eta} - \frac{1}{\rho}\right) s_k\right]^{-1} \frac{W}{z_k} \\ &= \left[\left(1 - \frac{1}{\rho}\right) - \left(\frac{1}{\eta} - \frac{1}{\rho}\right)\left(\frac{P_k}{P}\right)^{1-\rho}\right]^{-1} \frac{W}{z_k} \end{aligned}$$

As a consequence the equilibrium price vector is the solution to a set of $K+1$ nonlinear equations composed of K fixed point conditions, together with the definition of the aggregate price index. Define $\Phi\left(\cdot\,; z_k, P\right) : x \longrightarrow \frac{W}{z_k} \frac{1}{\left[\left(1-\frac{1}{\rho}\right) - \left(\frac{1}{\eta}-\frac{1}{\rho}\right)\left(\frac{P}{x}\right)^{\rho-1}\right]}$ and $\Psi\left(\mathbf{P}\right) \equiv \left[\sum_{k=1}^{K} \left(\frac{1}{P_k}\right)^{\rho-1}\right]^{-\frac{1}{\rho}}$. The K response-price equations can each be written $P_k = \Phi\left(P_k; z_k, P\right)$ and the function $\Phi\left(\cdot\,; z_k, P\right)$ is strictly decreasing in its first argument and maps $\left(P\left(\frac{1}{\eta}\frac{\rho-\eta}{\rho-1}\right)^{\frac{1}{\rho-1}}, +\infty\right)$ into $\left(\frac{\rho}{\rho-1}\frac{W}{z_k}, \frac{\eta}{\eta-1}\frac{W}{z_k}\right)$.

Lemma 2 (Price Index)**.** *The price index is i) more elastic with respect to the pricing decisions of the larger firms and ii) its variations are bounded above by the largest variations in prices.*

Proof: Differenciating the definition of the price index yields

$$\frac{dP}{P} = \left(\frac{\rho - 1}{\rho}\right) \sum_k s_k . \frac{dP_k}{P_k}$$

and therefore, the elasticity of the price index with respect to a change in a firm's individual price k is

$$\frac{dP/P}{dP_k/P_k} = \left(\frac{\rho - 1}{\rho}\right) s_k$$

Moreover, the changes in the price index are bounded above by the largest market-share weighted proportional change in firms'prices, i.e.

$$\left|\frac{dP}{P}\right| = \frac{\rho - 1}{\rho} \max_k \left|\frac{dP_k}{P_k}\right|. \tag{2.27}$$

where, by assumption 2, we have $\frac{\rho-1}{\rho} < 1$.

Lemma 3 (Prices elasticities with respect to the price index)**.** *The price elasticities of individual firms with respect to the price index are i) positive for all firms, ii) strictly smaller than one, and iii) increasing with the size of the firm, as measured by its market share.*

Proof: Differentiate the price response equation, and obtain

$$\frac{dP_k}{P_k} = \frac{x_k}{1 + x_k} \frac{dP}{P}$$

where $x_k \equiv (\rho - 1)\left(\frac{1}{\eta} - \frac{1}{\rho}\right) s_k \cdot \mu_k > 0$, where μ_k is the markup charged by firm k. To prove the second part, recall that the markups are increasing in the market share, that P_k is decreasing in the market share and that the function $x \to \frac{x}{1+x}$ is strictly increasing and mapping $[0, +\infty)$ onto $[0, 1)$.

Price impact of a productivity shock Proof: the change in firm k's price in response to a small productivity shock, can be decomposed as follows

$$d \ln P_k = d \ln \mu_k + d \ln W - d \ln z_k$$

where $\mu_k \equiv \frac{\epsilon_k}{\epsilon_k - 1}$ is the markup charged by firm k, W the market wage and z_k the idiosyncratic productivity of firm k. Now recall that in equilibrium the elasticity is linked to the market

share by the following relationship

$$\frac{1}{\epsilon_k} = \frac{1}{\rho}(1 - s_k) + \frac{1}{\eta} s_k$$

hence the markup is

$$\mu_k = \left[1 - \frac{1}{\rho} - \left(\frac{1}{\eta} - \frac{1}{\rho}\right) s_k\right]^{-1} \tag{2.28}$$

or alternatively, $\mu_k = \left[\frac{1}{\mu_\rho}(1 - s_k) + \frac{1}{\mu_\eta} s_k\right]^{-1}$, that is the markup μ_k is a weighted harmonic mean of the within and between markups. The change in markup is linked to the change in market share as follows

$$d \ln \mu_k = \mu_k \left(\frac{1}{\eta} - \frac{1}{\rho}\right) ds_k \tag{2.29}$$

Recall that the firm's market share's relationship with prices $s_k = \frac{(1/P_k)^{\rho-1}}{\sum_{j=1}^{K}(1/P_j)^{\rho-1}}$. This market share is affected by the price changes as follows:

$$\begin{aligned} ds_k &= -(\rho - 1)\frac{dP_k}{P_k} s_k \\ &+ (\rho - 1)\frac{dP_k}{P_k} s_k^2 \\ &+ (\rho - 1)\sum_{j \neq k} \frac{dP_j}{P_j} s_k \cdot s_j \end{aligned}$$

At the first order, firm k's market share responds to its own price change with an elasticity of $(\rho - 1)$. It increases when other firms raise their prices. Let $\mathbf{\Omega}$ be the matrix defined by

$$\mathbf{\Omega}_{kj} = \begin{cases} -\omega \mu_k s_k (1 - s_k) & j = k \\ \omega \mu_k s_k \cdot s_j & j \neq k \end{cases} \tag{2.30}$$

where $\omega \equiv \left(\frac{1}{\eta} - \frac{1}{\rho}\right)(\rho - 1)$ Then, the system of price responses to firms' idiosyncratic productivities shocks can be written compactly

$$d \ln \mathbf{P} = \mathbf{\Omega} d \ln \mathbf{P} - d \ln \mathbf{z}$$

and consequently, as long as $I - \Omega$ is nonsingular the price vector solve the following system

$$d \ln \mathbf{P} = -\left[I - \mathbf{\Omega}\right]^{-1} d \ln \mathbf{z} \tag{2.31}$$

Lemma 4. *The pass-through of productivity to prices is imperfect. In particular, we have*

$$T \geq 0, \quad \|T\|_\infty \leq 1, \quad \rho(T) < 1$$

where $T \equiv [I - \mathbf{\Omega}]^{-1}$ *denotes the pass-through matrix in this sector.*

Proof: $I - \mathbf{\Omega}$ is a positive stable P-matrix since it has positive diagonal entries and is row-diagonally dominant, i.e. $\|1 - \mathbf{\Omega}_{kk}\| > \sum_{j \neq k} \mathbf{\Omega}_{kj}$ for all k. It is therefore an M-matrix. Consequently, its inverse is well-defined and is a positive matrix The second part of the lemma follows from an application of the Ahlberg-Nilson-Varah bound $\frac{1}{\min_k \{|1-\mathbf{\Omega}_{kk}| - r_k(I-\mathbf{\Omega})\}} < 1$, where $r_k(I - \mathbf{\Omega})$ is the sum of the absolute values of the off-diagonal entries on row k. Axelsson and Kolotilina show that this bound is sharp for M-matrices.

2.C.2 Collusion

In this section we formally derive the main properties of the collusive industry equilibrium.

Proof of Proposition 1: Markups under collusion Consider a cartel in industry i composed of two firms, k_1 and k_2 . Let y denote the industry equilibrium output. We drop the subscripts i for sake of notation clarity. The objective function of firm k_1 is

$$\tilde{\Pi}(q_{k_1}, q_{k_2}) = \Pi_{k_1}(q_{k_1}, q_{k_2}) + \kappa \cdot \Pi_{k_2}(q_{k_1}, q_{k_2})$$

subject to the inverse demands $\frac{\tilde{P}_k}{P} = \left(\frac{q_k}{y}\right)^{-\frac{1}{\rho}} \cdot \left(\frac{y}{c}\right)^{-\frac{1}{\eta}}$. The profits of firm k_1 are $\Pi_{k_1} = \tilde{P}_k \cdot q_{k_1} - \frac{W}{z_{k_1}} \cdot q_{k_1}$. Consider the first order condition in q_{k_1}, we have

$$\begin{aligned} \frac{\partial \tilde{\Pi}(q_{k_1}, q_{k_2})}{\partial q_{k_1}} &= \frac{\partial \Pi_{k_1}(q_{k_1}, q_{k_2})}{\partial q_{k_1}} + \kappa \cdot \frac{\partial \Pi_{k_2}(q_{k_1}, q_{k_2})}{\partial q_{k_1}} \\ &= \left[1 - \left\{\frac{1}{\rho} + \left(\frac{1}{\eta} - \frac{1}{\rho}\right) \cdot s_{k_1}\right\}\right] P_{k_1} - \frac{W}{z_{k_1}} + \kappa \cdot \frac{\partial P_{k_2}}{\partial q_{k_1}} \cdot q_{k_2} \end{aligned}$$

The first two terms are exactly the same as in the FOC without collusion while the last term is the additional term created by the collusion, whereby a firm internalize only partially

the positive externality on the other members of the cartels.

$$\begin{aligned}\frac{\partial P_{k_2}}{\partial q_{k_1}}\cdot q_{k_2} &= \left[\left(\frac{1}{\rho}-\frac{1}{\eta}\right)\cdot P_{k_2}\cdot\left(\frac{q_{k_1}}{y}\right)^{-\frac{1}{\rho}}\cdot y^{-1}\right]\cdot q_{k_2}\\ &= \left(\frac{1}{\rho}-\frac{1}{\eta}\right)\cdot P_{k_2}\cdot q_{k_2}^{\frac{1}{\rho}}q_{k_1}^{\frac{-1}{\rho}}\cdot\left(\frac{q_{k_2}}{y}\right)^{1-\frac{1}{\rho}}\\ &= \left(\frac{1}{\rho}-\frac{1}{\eta}\right)\cdot P_{k_1}\cdot s_{k_2}\end{aligned}$$

Hence the result. The parameter $\kappa \in [0,1]$ controls the degree of symmetry of the cartel agreement. If $\kappa = 1$ then a member of the cartel cares equally about her own-profits than that of other members of the cartel. In this extreme case, all the members of the cartels set the same markups, that depends only on the sum of the equilibrium market shares of the cartel members. Conversely, $\kappa = 0$ corresponds to the baseline monopolistic competition case. The proof naturally extends to the case of cartels of an arbitrary size.

Proof:Pass-through under collusion. We can generalize the expression of the pass-through of productivity shocks on prices in the presence of a cartel, the price responses solve the following system

$$d\ln\mathbf{P} = -\left(1-\mathbf{\Omega}\left[\kappa\right]\right)^{-1}d\ln\mathbf{z} \tag{2.32}$$

where

$$\mathbf{\Omega}_{kj}\left[\kappa\right] = \begin{cases}\mathbf{\Omega}_{kj} & k\notin\mathcal{C}\\ \mathbf{\Omega}_{kj}+\kappa\sum_{i\in\mathcal{C}\setminus\{k\}}\left(\frac{\mu_k}{\mu_i}\right)\mathbf{\Omega}_{ij} & k\in\mathcal{C}\end{cases}$$

The pass-through is unchanged for the firms not part of the cartel (rows $k \notin \mathcal{C}$) while there are extra terms for the firms in the cartels (rows $k \in \mathcal{C}$). These extra terms are the consequences of cartel members partially internalizing the price effects that they have on each other. The extra term is therefore composed of the sum of the pass-through on the other cartel members, rescaled by their relative markups and the collusion intensity.

$$d\ln\mu_k = \mu_k\left(\frac{1}{\eta}-\frac{1}{\rho}\right)\left[ds_k+\kappa\cdot\sum_{j\in\mathcal{C}\setminus\{k\}}ds_j\right]$$

Dampened Pass-through The pass-through of productivity gains to prices is smaller in the presence of a cartel. In particular, we have

$$0 < \|\mathbf{T}[\kappa]\|_\infty < \|\mathbf{T}\|_\infty < 1$$

where

$$\mathbf{T}[\kappa] \equiv [\mathbf{I} - \mathbf{\Omega}(\kappa])^{-1}$$

. PROOF: Apply again the ANV bound. It is actually a minimum, since we have an M matrices and notice that

$$a_{ii}[\kappa] > a_{ii}$$

and $r_i(A[\kappa]) < r_i(A)$ where $A[\kappa] = I - \Omega[\kappa]$.

Proofs of Proposition 2. Prices in cartelized industry. 1. Using the previous notation, the price response to a cartel of intensity κ is

$$d\ln \mathbf{P} = (1 - \mathbf{\Omega}[\kappa])^{-1} \cdot \vec{\kappa} \tag{2.33}$$

where

$$\vec{\kappa} = \begin{cases} \kappa\mu_k \left(\frac{1}{\eta} - \frac{1}{\rho}\right) \sum_{i \in \mathcal{C}\setminus\{k\}} s_i & k \in \mathcal{C} \\ 0 & k \notin \mathcal{C} \end{cases}$$

is the vector with the scalar κ on each row $k \in \mathcal{C}$ and 0 otherwise. Therefore cartels operate just as negative idiosyncratic technological shocks. 4. From the nested CES structure we have $\frac{d\ln y_i}{d\ln P_i} = -\eta$. Hence, when the sector's price index go up, the share of this sector in total consumption goes down.

2.C.3 Alternative Market Structure: Bertrand Competition

We can alternatively solve the model under the assumption that firms engage in a static game of Bertrand Competition. In the baseline case, the markups are

$$\epsilon(s) = \rho + (\eta - \rho) \cdot s$$

and

$$\epsilon(s) = \rho + (\eta - \rho) \cdot \left[s_k + \kappa \cdot \sum_{j \in \mathcal{C} \setminus \{k\}} s_j \right]$$

for members of cartel $\mathcal{C}$. We obtain qualitatively similar effects but slightly different magnitudes. Because the firm-specific elasticities are now arithmetic means instead of harmonic means, they are at least as large as in the Cournot case. And therefore the markups in the Bertrand setting are smaller than in the Cournot setting.

2.D Additional Tables

Table 2.6: Robustness: Alternative Timing

Dependent Variable	Market Share$_{kt}$		
Timing	5 Years Before Investigation	2 Years Before Sanction	2 Years Before Investigation
	(1)	(2)	(3)
AntiComp$_{it}$	-0.007*	-0.005*	-0.007**
	(0.004)	(0.002)	(0.0035)
AntiComp$_{it}$ × Colluder$_{kt}$	2.973***	4.391***	3.838**
	(0.937)	(1.573)	(1.498)
Revenue TFP$_{kt}$	0.154**	0.154**	0.154**
	(0.063)	(0.063)	(0.063)
2-digit Sector × Year FE	✓	✓	✓
5-digit Industry FE	✓	✓	✓
# Obs.	15,040,836	15,040,836	15,040,836

Notes: Standard errors clustered at the 2-digit sector level. *** p<0.01, ** p<0.05, * p<0.1.

Table 2.7: Results when Colluders are the Parent Firms

Dependent Variable	Market Share$_{kt}$				
	(1)	(2)	(3)	(4)	(5)
AntiComp$_{it}$	-0.06***	-0.053***	-0.0007*	-0.0055	-0.0054*
	(0.018)	(0.015)	(0.003232)	(0.0027)	(0.003)
AntiComp$_{it}$ × Colluder$_{kt}$	4.156**	3.366**	3.786**	3.373**	5.67**
	(1.66)	(1.60)	(1.56)	(1.599)	(2.525)
Revenue TFP$_{kt}$		0.055		0.15**	
		(0.014)		(0.063)	
# Colluders$_{it}$					-0.001
					(0.0014)
# Colluders$_{it}$× Colluder$_{kt}$					-0.323
					(0.191)
2-digit Sector × Year FE	✓	✓	✓	✓	✓
5-digit Industry FE			✓	✓	✓
Firm Size Controls		✓			
# Obs.	20,167,666	15,040,842	20,167,662	15,040,836	20,167,662

Notes: Standard errors clustered at the 2-digit sector level. *** p<0.01, ** p<0.05, * p<0.1. Firm size controls include the quantity of workers, capital and materials (in log).

2.E A Theory of Mergers

Our framework naturally extends into a theory of mergers and acquisition. This section sketches such as theory and characterizes the shape of the synergies required for mergers to be welfare-improving given the current market structure. We refer to the companion

paper for a fully-fledged theory and a quantitative exercise using French mergers over the last decades.

2.E.1 Market Power v. Efficiency

The formation of cartels and merging behavior are inter-related. For instance, Evenett, Levenstein, and Suslow (2001, p. 1245):) observe that cartel breakdowns are often followed by mergers."Vigilance should not end with a cartels punishment, as former price-fixers often try to effectively restore the status quo ante by merging or by taking other steps that lessen competitive pressures and raise prices'. This pattern seems even more prevalent in Europe. About 45 of percent of cartels reported by the European Commission between 2001 and 2010 were followed by mergers involving at least one of the former cartel members (Kumar et al. 2015).

While plain collusions are unambiguously harmful, the welfare implications of mergers are theoretically ambiguous. This ambiguity follows from the classic trade-off between efficiency and market power emphasized by Williamson 1968: i) the increase in the (combined) market shares, taken into consideration while setting the prices, raises the markups, while ii) the productivity gains from potential synergies can reduce the marginal costs of the firms, therefore exerting a downward pressure on prices. To embedd this tradeoff in our framework, let us consider these two countervailing forces in turn.

The first channel – the market power effect – is the main force studied in our model of collusion. Little change is therefore needed, except perhaps a re-interpration of our micro-foundations with a greater emphasis on control instead of financial ownership.

To assess the impact of the second channel, consider two firms k_1 and k_2 in the same sector i that decide to merge. Without loss of generality, we suppose that firms are indexed in decreasing order of their productivities and therefore $z_{ik_1} > z_{ik_2}$. After the merger is consumed, we suppose that, at least initially, the acquiring firms does not rebrand completely all the products of the firms it has absorbed (see Allain et al. 2017 for details about "slow rebranding" in the retail sector). Hence the number of firms in sector i does not change

but the maximization programs of the merging firms now take into account their common market share. The situation is analogous to situation of common ownership (see O'Brien and Salop (2000) or Schmalz et al). [23] The merging parties now each maximize their profits, internalizing the impact on the sister company. This mechanism is similar to the collusion case and therefore we have

$$\tilde{\epsilon}_i(k_1, k_2) = \left[\frac{1}{\rho}(1 - [s_{ik_1} + s_{ik_2}]) + \frac{1}{\eta}[s_{ik_1} + s_{ik_2}]\right]^{-1}$$

In the model, the merger perturbs the equilibrium prices through the two direct channels mentioned above: i) an increase in the markups of the target and the acquirer $\mu_{ik}, k \in \mathcal{M}$ and ii) a decrease in marginal costs of the merging firms i.e.the increase in $z_{ik}, k \in \mathcal{M}$. However, as emphasized in the collusion model in the main text, equilibrium prices are influenced by an additional indirect channel: the umbrella pricing effect, i.e. the equilibrium response of competitors to changes in the sector's price index. The equilibrium price response after the merger can therefore be written

$$d\ln \mathbf{P} = (\mathbf{I} - \mathbf{\Omega}[1])^{-1} \cdot [\vec{\boldsymbol{\kappa}} - d\ln \mathbf{z}] \tag{2.34}$$

where the vector $\vec{\boldsymbol{\kappa}}$ is defined as

$$\vec{\kappa} = \begin{cases} \mu_k \left(\frac{1}{\eta} - \frac{1}{\rho}\right) \cdot \sum_{i \in \mathcal{M}\setminus\{k\}} s_i & k \in \mathcal{M} \\ 0 & k \notin \mathcal{M} \end{cases}$$

The two vectors in the brackets represent the direct effects of the merger: the market power effect ($\vec{\boldsymbol{\kappa}}$) and the productivity effect ($d\ln \mathbf{z}$), while the matrix $(1 - \mathbf{\Omega}[1])^{-1}$ embeds the general equilibrium effects generated by the merger.[24]

In this static framework and in the absence of upstream or downstream effects, the welfare effects of the mergers are fully captured by the movements of price index in the industry. If the merger raises the price index, then the welfare of the representative consumer is reduced.

[23] Alternatively, if one supose that the absorbed firm disappears then mergers would reduce welfare somewhat mechanistically by taking a variety off the market.

[24] We have assumed that the merger is symmetric. Asymmetric mergers whereby the distribution of control after the merger is not balanced can also be analyzed in our framework, at the cost of extra notations.

Proposition 11 (Welfare-Improving Mergers)**.** *A merger is weakly welfare-improving if and only if*

$$\mathbf{S}^t \cdot (1 - \mathbf{\Omega}[1])^{-1} \cdot [\vec{\boldsymbol{\kappa}} - d\ln \mathbf{z}] \geq 0$$

where $\mathbf{S} = [s_i]$ *is the* $K \times 1$ *vector of the market shares.*

A sufficient condition for a merger to be welfare-improving is that the prices of both firms decrease after the merger is consumed. This is the case as long the efficiency gains are large enough.

Proposition 12 (Sufficient Condition for Welfare-Improving Mergers)**.** *A sufficient condition for a merger* $\mathcal{M}$ *to be welfare improving is*

$$d\ln z_k \geq \frac{\sum_{i\in\mathcal{M}\setminus\{k\}} s_i}{\eta\frac{\rho-1}{\rho-\eta} - s_k},\ \forall i \in \mathcal{M}$$

Proof. Remark that $d\ln z_k \geq \mu_k \left(\frac{1}{\eta} - \frac{1}{\rho}\right) \sum_{i\in\mathcal{M}\setminus\{k\}} s_i$, and then substitute for μ_i

In order to make this criterion operational as a useful screen for merger reviews, we make further assumptions regarding the form of synergies that are created.

2.E.2 Merging Technologies

Most mergers involve only two parties, one acquiring and one target company. In this section we consider mergers involving two firms, a *target* firm, T and an acquiring firms, A. We assume that the acquiring firm is the one with the highest productivity, which yields a simpler version of the criterion.

Corollary 1. *A sufficient condition for a two-party merger to be welfare improving is*

$$d\ln z_T \geq \frac{s_A}{\eta\frac{\rho-1}{\rho-\eta} - s_T} \text{ and } d\ln z_A \geq \frac{s_T}{\eta\frac{\rho-1}{\rho-\eta} - s_A}$$

To analyze the trade-off in more details and deliver a sharper criterion, we examine several types of synergies [25]

[25]See David (2015) for a search-based model of mergers with a similar functional form for the post-merger productivity. David (2015) argues that they are positive assortative matching between merging partners and that the amount of synergies can be recovered from the joint size distribution of the merging parties.

1. Pure Technological Transfer Suppose that when two parties merge the productivity of the acquirer does not change while the target benefits from the acquirer's productivity, i.e. $d \ln z_T = \frac{z_A}{z_T} - 1$. Then using the fact that $z_k = \frac{W}{\mu_k P_k}$ and then the expressions linking the markups and the market share, one can express the left hand side only in terms of the market shares $\frac{1-\left(1-\frac{\mu_\rho}{\mu_\eta}\right)s_T}{1-\left(1-\frac{\mu_\rho}{\mu_\eta}\right)s_A}\left(\frac{s_A}{s_T}\right)^{\frac{1}{\rho-1}}$. Figure 3 shows the regions where this criterion is satisfied. The figure reveals that unless synergies also improve the acquirer's productivity, only very few of the merger would be welfare enhancing.

2. Additive gains Suppose that when two firms k_1 and k_2 merge, their resulting productivity is

$$m\left(z_{k_1}, z_{k_2}\right) = z_{k_1}^{\alpha} z_{k_2}^{\beta}$$

where $\alpha + \beta$ captures the level of synergies created.

2. Log linear improvements Suppose that when two firms k_1 and k_2 merge, their resulting productivity is

$$m\left(z_{k_1}, z_{k_2}\right) = z_{k_1}^{\alpha} z_{k_2}^{\beta}$$

where $\alpha + \beta$ captures the level of synergies created.

Relation with Critical Loss analysis Critical Loss Analysis is a common criteria used in Merger reviews.[26] This criterion serves to define markets.[27] In this article the definition of markets is guided by the available data and the markets examined in the decision. Because merging firms have successfully used that criteria to argue in court for a broader market definition that the one asserted by the government. It is useful to detail how our criteria can relate to Critical Loss analysis.

[26]For instance the guidelines in the United States: U.S. Dep't of Justice & Fed. Trade Comm'n, Horizontal Merger Guidelines (1992,revised1997), http://www.ftc.gov/bc/docs/horizmer.htm.

[27]see Farrell Shapiro footnote 2 The term "Critical Loss" was introduced by Barry Harris and Joseph Simons in Focusing Market Definition: How Much Substitution Is Necessary? 12 RES. L. & ECON. 207 (1989).

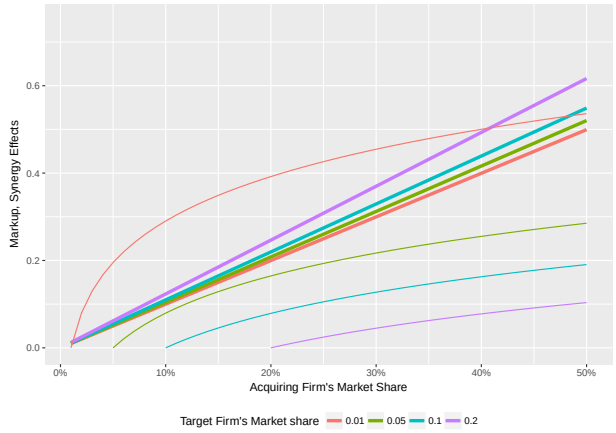


Figure 2.5: Merger tradeoff under pure technological transfer

The figure shows, for targets firms of different sizes (colors), the efficiency gains (concave thin lines) and the market power effects (bold, quasi-linear). Under this specification only mergers involving very small target companies are welfare enhancing.

Critical Loss analysis examines the counterfactual impact of a "significant and non-transitory increase in price" (SSNIP). The "Critical Loss" is defined as "the magnitude of lost sales that would (just) make it unprofitable for the hypothetical monopolist to impose a SSNIP, and compares it against the so-called Actual Loss of sales that would result from the SSNIP. If the Actual Loss, A, would be less than the Critical Loss, the SSNIP would be profitable". (Farrell Shapiro 2008). With a linear demand and constant marginal cost, Farrell and Shapiro show that the criterion can be written

$$A \geq \frac{s}{m+s}$$

where A is the counterfactual loss (confusingly called "actual loss" in the literature), s the market share and $m = \frac{p-c}{p}$ is the margin. It is possible to apply directly this criterion in our framework. The margin in our model is $m = \frac{1}{\epsilon(s)}$, hence the critical loss criterion becomes

$$A \geq \frac{s}{\frac{1}{\rho}(1-s) + \frac{1}{\eta}s + s}$$

CHAPTER 3

Active and Passive Antitrust Policies

How does antitrust policy shape business concentration and growth dynamics? I investigate the evolution of the firm size distribution in a macroeconomic model in which heterogeneous firms can grow organically or acquire competitors. I study the effects of antitrust policies under different market structures and with arbitrary initial distributions. In particular, I analyze the impact of threshold-based merger guidelines on concentration levels and productivity growth. I show that *passive* antitrust policies set the economy on an explosive path characterized by increasing size dispersion. Unless all mergers are blocked past a certain size, the size distribution of firms might not converge.

3.1 Introduction

Business concentration is growing in the United States and many European economies. Over the last three decades, the largest domestic firms of each sector command ever increasing market shares (Diez, Leigh, and Tambunlertchai 2018). A major concern with increased concentration is a decrease in competition, as measured by declining entry rates and increasing markups. Several hypothesis have been put forward to explain this rise in concentration. Grullon, Larkin, and Michaely 2015 highlights that the joint rise of markups and of concentration in the U.S. has coincided with a relaxation of antitrust standards (see also Gutiérrez and Philippon 2017). However there is limited causal evidence regarding the macroeconomic impact of antitrust policies. This paper provides a new theory of the firm size distribution in order to investigate the impact of antitrust on the firm-size distribution. To our knowledge, our model is the first to explicitly connect the stance of realistic antitrust policies with the evolution of macroeconomic quantities such as the economy's growth rate, moments of the firm size distribution, the level of markups, or the labor share.[1]

In our model, heterogeneous firms meet randomly and acquire competitors. Unless the proposed merger is blocked, the firm resulting form the merger has a higher productivity and more market power. While this leads to an increase in concentration, the net welfare effect is ambiguous. We consider several market structures giving rise to different markup distributions and, importantly, study the evolution of the firm size distribution, starting from any arbitrary firm-size distribution.

First, we take the search intensity as given, derive analytical expressions for the moments of the firm distribution, study the speed of convergence and the stationary properties of the economy. Then, we develop a tractable framework to analyze how antitrust policies shape the evolution of the firm size distribution. We study a large class of antitrust policy rules which are based on the sizes (or productivities) of the firms involved in the merger. Such threshold rules are the kind employed by competition authorities in most developed and

[1]For instance, antitrust policies in Edmond, Midrigan, and Xu 2018 are modeled as size-dependent taxes.

developing countries. We classify these rules as "passive" or "active" according to whether they lead to a diverging or stationary firm size distribution. A main result of our analysis is that, under threshold-based rules, unless the set of mergers that are approved is a compact set, the firm-size distribution does not converge. That is, even if an arbitrary small number of large mergers are cleared, the size distribution of firms diverges.

We then study how firms respond endogenously to antitrust policies by choosing their search intensity optimally and we show that, in equilibrium, the compactness condition derived for the exogenous case continues to hold. This is because as long as the gains from merging are strictly positive, large firms find it profitable to seek even very small firms to merge with. Unless entry becomes vanishingly small, this mechanism prevents the firm size distribution from converging. Because search incentives depend on the expected profits, we explore the interaction of antitrust policies with several types of market structures, resulting in different relationship between markup and size. Finally we study the consequences of antitrust policies on welfare and on the labor share, and study the design of optimal antitrust policies.

3.2 Literature Review

Firm size distribution: Ideas flows and Random growth This article connects several strands of the literature. First, modeling the business size distribution has a long history in macroeconomics. Lucas 1978 rationalized the business size distribution as stemming from managers' productivity distributed according to a Pareto distribution. Earlier, a first generation of random growth model (e.g. Kalecki 1945) derived convergence conditions for from simple statistical processes to produce stationary size distributions. A more recent literature has seized upon these ideas and emphasized two mechanisms at the heart of firm sizes' evolution: idea flows and innovation[2]. Our model of the evolution of firm sizes shares common features with these models and can accommodate some of the mechanisms present in the

[2]See Buera and Lucas 2018 for a survey on idea flows models and Luttmer 2010 for a survey on random growth models.

random growth literature. However, our merger and acquisition channel is fundamentally different. Contrary to Lucas Jr and Moll 2014 or Perla and Tonetti 2014, our growth mechanism does not require that the initial distribution has fat tails. In contrast, we show that even if the economy starts with a bounded distribution, sustained growth emerges endogenously from our merger mechanism. Even more drastically, sustain growth can emerge from an arbitrary mass point distribution of firms' initial productivities.

The random growth and idea flows literatures have largely abstracted from merger and acquisitions, with the notable exception of David (forthcoming). However, the core challenge in the design of merger policies is to balance potential synergies (productivity gains) with increases in market power, a trade-off absent from David's model, who considers potential productivity gains but shuts down any offsetting market power channel. His theory cannot serve as a basis of competition policy analysis. Moreover David does not investigate analytically the effects of antitrust policies on the dynamic evolution of the firm size distribution nor does he consider realistic antitrust policies.[3]

Another contribution of our paper is to derive analytical solutions to the partial differential equations governing the evolution of the firm sizes. We show how to use integral transforms to handle the convolution products that emerge naturally from merger behavior and derive closed-form solutions for arbitrary initial distribution. The mathematical techniques we use were introduced in economics by Duffie, Malamud, and Manso 2009 to study the percolation of information on OTC markets and explored further in a series of subsequent papers (Duffie, Giroux, and Manso 2010, Duffie, Malamud, and Manso 2014). Similar techniques, popular in physics to study Boltzman equations were also used by Chaney 2014 in his study of firms' international trade network formation.

Concentration and business dynamism A very active and recent literature has investigated the evolution business concentration and of firms' markups (see De Loecker et al 2019). The rise of concentration seems pervasive and has been connected a large number

[3]However David provides some useful simulations that help assess the importance of the merger channel. He founds that blocking all mergers would reduce the output by several GDP points.

of business dynamics patterns. Akcigit et al 2019 presents nine of these stylized facts: i) average markups have increased ii) the profit share of GDP has increased; iii) the labor share of output has gone down; iv) the rise in market concentration and the fall in labor share are positively associated; v) productivity dispersion of firms has risen, similarly, the labor productivity gap between frontier and laggard firms has widened.;vi) firm entry rate has declined. vii)the share of young firms in economic activity has declined. viii) job reallocation has slowed down. ix) the dispersion of firm growth has decreased. We show that some of these facts can be rationalized in our model.

Optimal antitrust rules The basic trade-off in the decision on whether to authorize horizontal mergers involves balancing the potential efficiency gains with the increased market power (Williamson 1968). Merger approval rules have been the focus of a small and recent theoretical literature started by Besanko and Spulber 1993. Other recent papers in this literature include Nocke and Whinston 2010, Ottaviani and Wickelgren 2011, and Neven and Röller 2005. More recently, Nocke and Whinston 2013 demonstrate that competition authorities should impose "a tougher standard on mergers involving larger merger partners (in terms of their premerger market share)". The set of active policies rules characterized in our paper has many similarities with theirs.

Empirical evidence on mergers Finally, a rather voluminous body of work in empirical industrial organizations has studied the impact of mergers, albeit usually focusing on very specific sectors such as airlines, banking, or the health care sector This specificity makes it difficult to draw general patterns that could be directly integrated in a macroeconomic model of the kind we are considering. Recent surveys, however, tend to show that mergers might have had some anti-competitive effects. For instance, according to Ashenfelter, Hosken, and Weinberg 2014: "the empirical record shows that mergers in oligopolistic markets can raise consumer prices". Regarding vertical mergers, Salop 2017 suggests that the current stance might be too lax.[4] The motivations for mergers, the stock market response to mergers,

[4]While our framework does not distinguish the upstream/downstream relationship between the merging firms, our preferred interpretation is that our model is more suited to the analysis of horizontal mergers.

and explanations for mergers to happen in waves have been studied in numerous finance articles (e.g. Maksimovic, Phillips, and Yang 2013). Perhaps closest to the spirit of our paper, Wollman (AERI 2018) studies the relaxation of merger pre-notification thresholds that happened in the U.S. in 2000. Wollman documents a "stealth concentration", finds a large deterrent effects of antitrust enforcement and estimate that the relaxation of antitrust policies "could account for 30% of the total change in four-firm revenue concentration" from 1994 to 2011.

Outline The outline of the paper is as follows. Section 2 lays out a simple model of the evolution of the firm size distribution when firms can grow by acquiring competitors. Search intensities and entry rates are taken as given. Section 3 introduces a framework to study antitrust policies and analyzes how these policies shape the dynamics of the economy. Section 4 analyzes equilibrium behavior when firms are chosing search intensities optimally. Section 5 studies welfare implications and the design optimal antitrust rules. Various extensions of the model are presented in Section 6. Mathematical proofs are relegated in the appendix.

3.3 Model

Transaction costs of many forms prevents the market for corporate control to be perfectly competitive. Instead, merging parties must overcome informational and financial frictions, just like in many OTC markets. Let $F(z,t)$ denote the cumulative measure of firms with productivity index at most equal to z at time t. For clarity, we sometimes use the more compact notation $F_t(z)$ for the cumulative density and f_t for the probability density.[5] The total mass of firms in the economy at time t is denoted $M_t := \lim_{z\to+\infty} F(z,t)$. Withour loss of generality, we assume that $M_0 = 1$.

[5]There are many sources of productivity differences across firms, including intangible capital or management.

3.3.1 Baseline Model

Firms meet randomly at a constant rate λ and search is undirected. When two firms meet, we assume that the acquiring firm is always the one with the largest productivity and the other one is the target.[6] After the meeting, the acquiring firm's productivity index increases and becomes the sum of the two productivity indices $m(z_1, z_2) = z_1 + z_2$. The target firm then leaves the economy. This simple additive merger technology is chosen for tractability. Random search is consistent with the fact that given an acquirer's size, there is substantial variance in the size of the acquired firms in the cross-section, while directed search would tend to have acquirer target firms with a specific size. While we formulate simple additive synergies in productivity, we investigate in Section 3 a much large class of synergies in terms of output and profits, that depends on the market structure and on the properties of the production function. The density of firms has the following evolution

$$\frac{\partial f(z,t)}{\partial t} = \underbrace{-\lambda f(z,t)}_{\text{Firm z is Acquirer or Target}} + \underbrace{\frac{\lambda}{2}\int_{-\infty}^{+\infty} f(y,t) f(z-y,t)\, dy}_{\text{Mergers resulting in type z Firm}} + \underbrace{\eta_t g(z,t)}_{\text{Entry}} \tag{3.1}$$

The first term on the right hand side is the flow of firms with productivity index z at time t whose productivity changes after a meeting (either because they are acquired or grow larger after merging with a target). The second term corresponds to the flow of mergers resulting in a firm with productivity z. That is, all the mergers involving pairs of firms productivity indices $(y, z-y)$ where $y \in [0, z]$. Because a mass $\lambda \int_z f(z,t)\, dz = \lambda M_t$ of firms merges every period, $\frac{\lambda}{2} M_t$ targets disappear from the firm size distribution.[7] After each merger, the target firm is removed and the measure of firms shrinks. In order to prevent the mass of firms from shrinking to zero, entry of new firms is required. Suppose that there is a fringe of firms with productivity measure $g(z,t)$ entering at rate η_t then the evolution of the mass of firms in this economy is $\frac{\partial}{\partial t} M_t = -\frac{\lambda}{2} M_t + \eta_t$. If there is no entry of new firms ($\eta_t = 0,\ \forall t$), then

[6] In practice, only a small fractions of mergers involve a smaller firm acquiring a larger one (see David, Wollman).

[7] See Duffie et al for details on the technical assumptions required for the LLN to apply in this context.

the mass of firms shrinks exponentially, and at any time is given by $M_t = M_0 e^{-\frac{\lambda}{2}t}$. Unless specified, we will focus on situations where the mass of firms in this economy is converging to some strictly positive number.

Because of the integral on its right hand side, the partial differential equation 3.1 is not linear. Fortunately, the analysis can be carried out by introducing the convolution operator $*$ and using integral transform techniques. These techniques are the standard tool in physics to study the behavior of gases following Boltzman equations. They have been popularized in economics recently in a series of paper by Duffie and coauthors to study the percolation of information between traders on OTC markets and by Chaney for the study of firms networks in international trade. The convolution product of any two measurable functions f_1 and f_2 is defined as $f_1 * f_2 = \int_{-\infty}^{+\infty} f_1(y) f(z-y)\, dy$ and we denote f^{*n} the n-fold convolution of f with itself.[8] Therefore equation 3.1 can be written:

$$\frac{\partial f(z,t)}{\partial t} = -\lambda f(z,t) + \frac{\lambda}{2} f(z,t) * f(z,t) + \eta g(z,t) \tag{3.2}$$

The Fourier transform[9] of a continuous function f is defined as $\mathcal{F}[f](s) = \int_{-\infty}^{+\infty} e^{-izs} f(z)\, dz$, for $s \in \mathbb{R}$. For any two measurable functions f_1 and f_2, an important properties of the Fourier transform is that $\mathcal{F}[f_1 * f_2] = \mathcal{F}[f_1] \times \mathcal{F}[f_2]$. We therefore take the Fourier transform on both sides of equation 3.2 and, denoting $\hat{f} = \mathcal{F}[f]$, we have

$$\frac{\partial \hat{f}(s,t)}{\partial t} = -\lambda \hat{f}(s,t) + \frac{\lambda}{2} \hat{f}^2(s,t) + \eta_t \hat{g}(s,t) \tag{3.3}$$

In order to analyze this equation, we make certain assumptions regarding the entry distribution and the entry rate. We investigate the properties of the model under alternative assumptions for the entry distribution in Section 6.

Assumption 1: $g_t = f_t$ **and** $\eta_t = \eta < \lambda$ Under these assumptions, entering firms draw from the current productivity distribution and the amount of entry is scaled by the current

[8] By convention, $f^{*1} = f$ and $f^{*2} = f * f$. n refers to the number of times f is written in the expression.

[9] Other integral transforms are suitable to handle the convolution product, for instance the Laplace transform defined as $\mathcal{L}[f](s) = \int_0^{+\infty} e^{-zs} f(z)\, dz$

mass of firm, yielding a constant entry rate. The firm distribution evolves as follows:

$$\frac{\partial \hat{f}(s,t)}{\partial t} = -(\lambda - \eta)\hat{f}(s,t) + \frac{\lambda}{2}\hat{f}^2(s,t) \tag{3.4}$$

For any $s \in \mathbb{R}$, this equation is a Ricatti equation with constant coefficients that admits a closed-form solution. The solution only depends on the search and entry parameters and moments of the initial distribution captured by its Fourier transform. We ensure that we have recovered the entire function $\hat{f}_t$ by checking continuity at every point. For any $s \in \mathbb{R}$, the solution of (3.4) is

$$\hat{f}_t(s) = \frac{\hat{f}_0(s)}{e^{\tilde{\lambda}t}\left[1 - \frac{1}{\tilde{\lambda}}\frac{\lambda}{2}\hat{f}_0(s)\right] + \frac{1}{\tilde{\lambda}}\frac{\lambda}{2}\hat{f}_0(s)} \tag{3.5}$$

where $\tilde{\lambda} = \lambda - \eta$ and $\hat{f}_0(s)$ is the Fourier transform of the arbitrary initial distribution of productivity $f_0(z)$. To obtain an analytical expression for f, we first expand the expression into a Wild summation:

$$\hat{f}(s,t) = e^{-\tilde{\lambda}t}\sum_{n\geq 1}\left[\frac{\lambda}{2}\frac{1}{\tilde{\lambda}}\left(1 - e^{-\tilde{\lambda}t}\right)\right]^{n-1}\hat{f}_0(s)^n \tag{3.6}$$

Fourier transforms are linear and under suitable conditions a function can be uniquely recovered from its Fourier transform.[10] Because taking the Fourier transform of the right hand side of the following equation yields the right hand side of equation (3.6)we conclude that the analytical expression for the solution f of (3.1) is

$$f_t(z) = \sum_{n\geq 1} e^{-\tilde{\lambda}t}\left[\frac{\lambda}{2}\frac{1}{\tilde{\lambda}}\left(1 - e^{-\tilde{\lambda}t}\right)\right]^{n-1} f_0(z)^{*n} \tag{3.7}$$

From this expression, it is clear that, at any point time, the density of firms with a given index z corresponds to a mixture of convolutions of the initial distributions. For a firm to reach productivity z, it needs to have either started at z and never merged, or started with a lower productivity and underwent a finite number of mergers n. Each terms of the infinite sum precisely count the probability for such a sequence of mergers to happen. while the We now proceed to study the evolution of the firm-size distribution when the number of firms stays constant.

[10] See Mathematical Appendix.

Assumption 2: $\eta = \frac{\lambda}{2}$ **.** As discussed previously, an active entry is necessary to prevent the mass of firms in the economy from shrinking to zero. Combined with Assumption 1, the condition that $\eta = \frac{\lambda}{2}$ guarantees that the mass of firms is constant because each acquired target is replaced by a new firm drawn from the current size distribution. The expression of the Fourier solution simplifies to

$$\hat{f}_t(s) = \frac{\hat{f}_0(s)}{e^{\frac{\lambda}{2}t}\left[1 - \hat{f}_0(s)\right] + \hat{f}_0(s)} \tag{3.8}$$

And similarly, we can simplify the expression for thecorresponding firm size distribution.

$$f_t(z) = e^{-\frac{\lambda}{2}t} \sum_{n \geq 1} \left(1 - e^{-\frac{\lambda}{2}t}\right)^{n-1} f_0(z)^{*n} \tag{3.9}$$

We now use both expressions to characterize the evolution of the moments of the firm size distribution starting from an arbitrary distribution f_0.

Characterizing the Moments of the distribution Integral transforms such as the Fourier transform and the Laplace transform, when applied to a probability distribution, are intimately linked to the characteristic function. The moments of any random variable X_t with probability distribution function f_t, can be computed from the successive derivatives of the Fourier transform: $\mathbb{E}[X^k] = (-i)^k \varphi_f^{(k)}(0) = i^k \hat{f}^{(k)}(0)$. Let $m_t^k \equiv \mathbb{E}[X^k]$ denote the k^{th} uncentered moment of the distribution.

Proposition 13. *For any initial distribution f_0, the mean of the firm size distribution f_t grows exponentially at rate $\frac{\lambda}{2}$, and in particular*

$$m_t^1 = m_0^1 e^{\frac{\lambda}{2}t} \tag{3.10}$$

where $m_t^1 \equiv \int z \cdot f_t(z)\, dz$.

This proposition demonstrates that our model is able to generate sustained long-run growth regardless of the size of the right-tail of the distribution. This result stands in stark contrast with the findings in Lucas Jr and Moll 2014 or Perla and Tonetti 2014, who require an initial distribution with fat tails. Actually, we do not even need a right hand tail: bounded

initial distributions also generate sustained growth. We show in Example 1 that growth is sustained in our model even if all the firms start at the same mass point. Naturally, a higher search intensity leads to higher long-run growth. We derive similarly the second moment of the distribution at any time t.

Proposition 14. *For any initial distribution f_0, the variance of the distribution grows asymptotivally at rate λ, and in particular*

$$m_t^2 - \left(m_t^1\right)^2 = \left(m_0^1\right)^2 e^{\lambda t} + e^{\frac{\lambda}{2}t}\left[m_0^2 - 2\left(m_0^1\right)^2\right] \tag{3.11}$$

where the first moment is $m_t^2 \equiv \int z^2 \cdot f_t(z)\, dz$.

The expression of the variance is composed of two terms. The first one grows at rate λ while the second one grows slower at rate $\frac{\lambda}{2}$, which is the same as the growth rate of the mean. Asymptotically, the variance depends only the mean of the initial distribution, not on its variance. In particular, even with an initial variance of 0, the variance will grow to infinity. We now illustrate the basic mechanics of the model on a few examples where the initial firm size distribution is specified and the cumulative density function can be derived in closed-form.

3.3.2 Examples

Initial Dirac mass $f_0(z) = \delta_{z_0}$. Suppose that the economy starts with a unit mass of firms, all with the same productivity z_0. Then at any time t the probability distribution function is

$$f_{z_0,t}(x) = \begin{cases} e^{-\frac{\lambda}{2}t}\left[1 - e^{-\frac{\lambda}{2}t}\right]^{n-1} & \text{if } \ z = n \cdot z_0 \\ 0 & \text{otherwise} \end{cases}$$

where we used that the initial distribution can be represented by a Dirac distribution δ_{z_0} and that the n-fold convolution of the diract distribution is ${\delta_{z_0}}^{*n} = \delta_{n \cdot z_0}$. , and the cdf is a step function

$$F_{z_0,t}(x) = \begin{cases} 0 & z < z_0 \\ 1 - \left[1 - e^{-\frac{\lambda}{2}t}\right]^n & \text{if } \ nz_0 \leq z < (n+1)\, z_0 \end{cases} \tag{3.12}$$

The distribution converges pointwise to 0 , i.e. $\forall x, \lim_{t\to\infty} F_{z_0,t}(x) = 0$, and therefore there is no steady state. For large values of t the cdf is approximately equal to $F_{z_0,t}(nz_0) \approx (n-1)\, e^{-\frac{\lambda}{2}t}$. The mean at time t is

$$m_t^1 = z_0 e^{\frac{\lambda}{2}t}$$

Exponential Distribution $f_0(z) = \alpha e^{-\alpha z}, z \geq 0$**.** Suppose that the initial distribution of productivities is exponential. By rearranging equation (3.6) we find that the solution is the exponential distribution with parameter $\alpha e^{-\frac{\lambda}{2}t}$

$$f_t(z) = \alpha e^{-\frac{\lambda}{2}t} \cdot e^{-\alpha e^{-\frac{\lambda}{2}t} z}$$

At any point in time the mean is $\alpha^{-1} e^{\frac{\lambda}{2}t}$ and the variance $\alpha^{-2} e^{\lambda t}$ and they both diverge to infinity. The distribution does not converge towards a steady state as the parameter of the distribution gets arbitrary close to 0.

3.4 Antitrust Policy

Competition Authorities in many OECD countries are entrusted of protecting competition. Their main tools are merger control and investigation of antitrust violations. In order to prevent excessive concentration, merger control is the tool of choice when the large firms can become too dominant in their market. Because investigating mergers is costly and time-consuming, merger control is very ineffective in the absence of mandatory pre-merger notification.[11] Blocking a merger afer it is consummated entails costly legal procedures (see Wollman AERI). As a consequence, pre-merger notifications play a key for antitrust enforcement. While antitrust regulations have been in place since the beginning of the century, it is only in 1977 that the US Congress passed the Hart-Scott Rodino Improvement

[11]One could write a formal model of investigation decisions. In the absence of pre-merger notifications, information is costly to acquire for the regulator, and its resources are limited. Because it is likely less expensive for the firms to provide the required information than for the regulator to acquire it, in the absence of pre-merger notifications the regulator would have much less information in equilibrium.

Act, a significant amendment to antitrust policies, which mandates firms to notify mergers above a certain size criterion. The HSR criteria are explicitely based on the size of the targeted firm[12].

After the notification, the FTC decides whether to clear or block he merger.[cite figures] How these antitrust decisions are taken in practice has changed over time, as detailed in the successive FTC Horizontal Merger Guidelines. In the U.S., mergers that result in a combined concentration exceeding a high HHI threshold used to be presumed anti-competitive and were automatically blocked. Such a *per se* rule has prevailed until the early 1980's. Antitrust standards have moved away from the *per se rule* and shifted to the so-called "rule of reason" where no merger is to be rejected *ex ante* but evaluated on its own merits. Instead, each case is decided in a discretionary fashion. The 2010 US horizontal mergers guidelines explicitely spell out a set of HHI thresholds that serves as a screen.[13] These considerations motivate the following definitions and formulations of anitrust policy rules.

3.4.1 Active and Passive Policies

We now introduce antitrust policy rules in our model. Antitrus policy rules are defined as mappings that associate to each pair of firms, indexed by their productivities, a probability. We require that these policies satisfy certain "consistency" conditions.

Definition 3.1. *An antitrust policy rule is a function* $p : \mathcal{Z}^2 \to [0, 1]$ *such that:*

[12]"a filing is required when both a "size-of-persons" and "size-of-transaction" test are met. The first of these tests is met if (a) the acquired party is engaged in manufacturing, has sales or assets of $10 million or more, and the acquiring party has sales or assets of $100 million or more, (b) the acquired party is not engaged in manufacturing, has assets of $10 million or more, and the acquiring party has sales or assets of $100 million or more, or (c) the acquired party has total assets of $10 million or more and the acquiring party has sales or assets of $100 million or more. The second of these tests is met if (a) 15% or more of the voting securities or assets are acquired or (b) an aggregate amount of voting securities or assets in excess of $15 million are acquired" Wollman AERI 2019.

[13]"The purpose of these thresholds is not to provide a rigid screen to separate competitively benign mergers from anticompetitive ones, although high levels of concentration do raise concerns. Rather, they provide one way to identify some mergers unlikely to raise competitive concerns and some others for which it is particularly important to examine whether other competitive factors confirm, reinforce, or counteract the potentially harmful effects of increased concentration. *The higher the post-merger HHI* and the increase in the HHI, the greater are the Agencies' potential competitive concerns and *the greater is the likelihood* that the Agencies will request additional information to conduct their analysis." (empasis mine) FTC Horizontal Merger Guidelines 2010.

i) p is symmetric, i.e. $p(x,y) = p(y,x)$ for any $x, y \in \mathcal{Z}$; and

ii) p is weakly increasing, i.e. for any $x, x', y, y' \in \mathcal{Z}$ such that $x \leq x'$ and $y < y'$ we have $p(x,y) \geq p(x',y')$.

Let $\mathcal{R}(\mathcal{Z}^2)$ denote the space of antitrust policy rules. Our "consistency" requirement means that under a policy rule $p \in \mathcal{R}(\mathcal{Z}^2)$ any merger between two firms with produtivities x and y is accepted with probability $p(x,y)$, regardless of whether x is attempting to acquire y or y is attempting to acquire x. We also require that mergers between smaller firms are approved with greater probability than mergers between large firms, which is essentially what is implied by the FTC guidelines. Note that these policies can be discontinuous. Under an antitrust policy p, the firm size distribution evolves as follows:

$$\frac{\partial f_t(z)}{\partial t} = \underbrace{\eta f_t(z)}_{\text{Entry}} \underbrace{-\lambda \left[\int_0^{+\infty} p(y,z) f_t(y)\, dy \right] \cdot f(z,t)}_{\text{Approved Deals with firms z as Acquirer or Target}} + \underbrace{\frac{\lambda}{2} \int_0^{+\infty} p(z-y,y) f_t(y) f_t(z-y)\, dy}_{\text{Approved Mergers resulting in type z firm}} \tag{3.13}$$

We now introduce a partial ordering on the space $\mathcal{R}(\mathcal{Z}^2)$.

Definition 3.2. *An antitrust policy p is more lax than a policy p' , denoted $p \overset{\ell}{\succeq} p'$ if and only $\forall (x,y), p(x,y) \geq p'(x,y)$.*

In other words, a policy p' is more lax than a given policy p if any merger has a higher probability of being cleared under p' than under p. The order is partial since not all policies can be ranked, for instance neither $p : (x,y) \mapsto \frac{1}{2}$ nor $p' : (x,y) \mapsto e^{-(x+y)}$ is laxer than the other one.

Definition 3.3. *An antitrust policy rule p is active if there exists an entry rate $\eta_p > 0$ and a non-degenerate distribution f_∞ such that given the initial firm size distribution f_0, f_∞ is a stationary solution to Kolmogorov equation (3.13) when $\eta = \eta_p$.*

We denote by $\mathcal{P}_{f_0}(\lambda)$ the set of *passive* policy rules and $\mathcal{A}_{f_0}(\lambda)$ the set of *active* policy rules. Our characterization of antitrust policies as "active" or "passive" echoes Taylor's

characterization of monetary policies (see Taylor). "Active" antitrust policies control the evolution of the moments of the firms size distribution just like active monetary policy rules prevent prices from blowing up. Conversely, like a passive monetary policy leads the economy on an explosive path of inflation, a passive antitrust policy rule leads the economy on an explosive path. We now define a special class of policy rules that is particularly helpful to analyze the evolution of the firm size distribution.

Definition 3.4. *A threshold-based antitrust policy rule p is characterized by a finite set of pairs $\{(\tau_i, \phi_i)\}_{i \leq N_i}$, where $\tau_i \in \mathcal{Z}$ is a productivity threshold and $\phi_i \in [0,1]$ is the probability that the merger is approved, such that any merger between any two firms with produtivities x and y is accepted with probability*

$$p(x,y) = \begin{cases} 1 & \text{if } \ 0 \leq \min\{x,y\} \leq \tau_1 \\ \phi_i & \text{if } \ \tau_i \leq \min\{x,y\} < \tau_{i+1} \end{cases}$$

Let $\mathcal{R}_{\mathcal{T}}(\mathcal{Z}^{\mathbf{2}}) \subset \mathcal{R}(\mathcal{Z}^2)$ denote the set of threshold-based policy rules and $\mathcal{R}_{\mathcal{T}_n}(\mathcal{Z}^2)$ the set of threshold-based policy rules with at most n thresholds, such that $\mathcal{R}_{\mathcal{T}_1}(\mathcal{Z}^2) \subset \mathcal{R}_{\mathcal{T}_2}(\mathcal{Z}^2) \subset ... \subset \mathcal{R}_{\mathcal{T}_n}(\mathcal{Z}^2) \subset ... \subset \mathcal{R}_{\mathcal{T}}(\mathcal{Z}^2) = \cup_n \mathcal{R}_{\mathcal{T}_n}(\mathcal{Z}^2)$. The threshold-based antitrust policies in our model offer a good description of the actual implementation of antitrust *pre-merger notification* policies around the globe as explained above. However, whether a merger is blocked or approved can depend on a further set of considerations and the results of an in-depth investigation. Until the 1980's large mergers in the U.S. were deemed illegal *per se* and blocked. One interpretation of our threshold-based rules is therefore a combination of a pre-merger notification threshold and a *per se* decision rule.

3.4.2 Firm size distribution under a threshold-based antitrust policy

We now analyze the impact of introducing a simple antitrust policy with only one threshold in the baseline model of section 2. The evolution of the firm size distribution can be broken down in three regions depending on the size of the firm. For a small firm with $z < \tau$, the

KFE is

$$\frac{\partial f(z,t)}{\partial t} = \underbrace{\eta f(z,t)}_{\text{Entry}} - \underbrace{\lambda f(z,t)}_{\text{Mergers outflows all accepted}} + \underbrace{\frac{\lambda}{2} f * f(z,t)}_{\text{All Mergers inflows accepted}} \tag{3.14}$$

since all the mergers involving this firm are allowed because the target size is at most $z < \tau$ while mergers forming an entity of size z must combine firms below the target, leaving the law of motion is unchanged on $[\underline{z}, \tau]$. For a firm of intermediate size, i.e. with $\tau < z < 2\tau$, the KFE is

$$\frac{\partial f_t(z)}{\partial t} = \underbrace{\eta f_t(z)}_{\text{Entry}} - \underbrace{\lambda F_t(\tau) f_t(z)}_{\text{Unscrutinized mergers outflows}} - \underbrace{\lambda \phi [1 - F_t(\tau)] f_t(z)}_{\text{Scrutinized mergers outflows}} + \underbrace{\frac{\lambda}{2} f_t * f_t(z)}_{\text{Deals reaching } z \text{ not scrutinized}} \tag{3.15}$$

because the mergers outflows involving targets smaller than $\tau$are not scrutinized. All the mergers inflows have at least one firm with productivy strictly smalle than τ, and as a consequence the target firm must be below the threshold and the deal is not scrutinized. Finally, for large firms with $z > 2\tau$,

$$\begin{aligned} \frac{\partial f_t(z)}{\partial t} &= \underbrace{\eta f_t(z)}_{\text{entry}} - \underbrace{\lambda F_t(\tau) f_t(z)}_{\text{Unscrutinized merger outflows}} - \underbrace{\lambda \phi [1 - F_t(\tau)] f_t(z)}_{\text{Scrutinized merger outflows}} \\ &+ \underbrace{\frac{\lambda}{2} \int_{-\infty}^{\tau} f_t(y) f_t(z-y)\, dy + \frac{\lambda}{2} \int_{z-\tau}^{+\infty} f_t(y) f_t(z-y)\, dy}_{\text{Unscrutinized merger inflows}} \\ &+ \underbrace{\frac{\lambda}{2} \phi \int_{\tau}^{z-\tau} f_t(y) f_t(z-y)\, dy}_{\text{Scrutinized merger outflows}} \end{aligned} \tag{3.16}$$

since only the mergers involving small targets are left unscrutinized. The two terms on the second row can be combined usuing the symmetry of the convolution product and the KFE simplifies into

$$\begin{aligned} \frac{\partial f_t(z)}{\partial t} &= \eta f_t(z) - \lambda [F_t(\tau) + \phi [1 - F_t(\tau)]] f_t(z) \\ &+ \lambda \int_{z-\tau}^{\infty} f_t(y) f_t(z-y)\, dy + \tfrac{\lambda}{2} \phi \int_{\tau}^{z-\tau} f_t(y) f_t(z-y)\, dy \end{aligned} \tag{3.17}$$

These diffential equations completely determine the evolution of the firm size distribution under a threshold-based antitrust policy. The behavior of these equations change for different values of the size thresholds $\tau$and the blocking rate ϕ, which allows us to characterize active and passive policies.

Proposition 15. *The set of active policy rules, $\mathcal{A}_{f_0}(\lambda)$, is not empty.*

Proof: set $\tau = 0$ and $\phi = 0$. Then all mergers are scrutinized and none is allowed, therefore the distribution is stationary.

More generally, with $\phi = 0$, large firms can only grow by merging with very small firms, i.e. the ones below the threshold. Intutitively, when all these firms have been acquired, M&A activity stops. In fact, because of entry, the stock of these small firms is partly replenished at each period. Section 2 shows that if the entry distribution tracks the current distribution this process leads the firm size distribtuion to shift to the right indefinitely such that $\lim_{t\to\infty} F_t(\tau) = 0$.

Proposition 16. *If $p \in \mathcal{A}_{f_0}(\lambda)$ then for any p' such that $p \overset{\ell}{\succeq} p'$, $p' \in \mathcal{A}_{f_0}(\lambda)$.*

Passivity of antitrust policies is a transitive property in the sense that any policy more lax than a passive policy is also a passive policy. The converse is also true for active policies. The main consequence of this result is that to prove that a given policy is active, it is sufficient to find a more lax policy that can be proved to be active. Because of the simple structure of threshold-based, we can analyze these policies and derive conclusions for more complicated policies.

A first important result of this paper is a sufficient condition for active policy rules. Combined with Proposition 3, it guarantees that all the policy rules with finite support are active.

Proposition 17. *For any $p \in \mathcal{R}_{\mathcal{T}_1}(\mathcal{Z}^2)$, if $\phi = 0$ and $\lambda > 0$ then $p \in \mathcal{A}_{f_0}(\lambda)$.*

Proof:

If $\phi = 0$ then consider the law of motion for a firm of size $z_t > \tau$

$$\frac{\partial f_t(z)}{\partial t} = \eta f_t(z) - \lambda F_t(\tau) f_t(z) + \lambda \int_0^\tau f_t(y) f_t(z-y)\, dy \tag{3.18}$$

Now looking at the law of motion for all the $z < \tau$ it is clear from section 2 that if $\lambda > 0$ then $\lim_{t\to\infty} F_t(\tau) = 0$ and therefore mergers stop above the threshold. By using the fact that,

by definition of the cdf, given t, F_t is increasing then F_t converge for any $z \in \mathcal{Z}$. This result extends naturally to any threshold-based policy with an arbitrary large but finite number of thresholds.

Corollary. *For any $N \in \mathbb{N}$ and $p \in \mathcal{R}_{\mathcal{T}_N}(\mathcal{Z}^2)$, if $\phi_N = 0$ and $\lambda > 0$ then $p \in \mathcal{A}_{f_0}(\lambda)$.*

Therefore a very large class of active antitrust policies can be analyzed since many complicated policy rules can be tailored by selecting enough thresholds and approving rates. We now prove an opposite result: if the blocking rate above the highest size threshold is not set to 0, then the firm size distribution does not converge.

Proposition 18. *For any $p \in \mathcal{R}_{\mathcal{T}_1}(\mathcal{Z}^2)$, if $\phi > 0$ and $\lambda > 0$ then $p \in \mathcal{P}_{f_0}(\lambda)$.*

Proof: If $\phi > 0$ we have

$$\begin{aligned} \frac{\partial f_t(z)}{\partial t} &= \eta f_t(z) - \lambda \left[F_t(\tau) + \phi \left[1 - F_t(\tau)\right]\right] f_t(z) \\ &+ \lambda \int_{z-\tau}^{\infty} f_t(y) f_t(z-y)\, dy + \tfrac{\lambda}{2}\phi \int_{\tau}^{z-\tau} f_t(y) f_t(z-y)\, dy \end{aligned} \tag{3.19}$$

Towards to a contradiction, suppose that there exist a entry rate $\eta > 0$ for which there is a nondegenerate stationary solution f, we show that this equation cannot hold for all z. There are two possible cases, $F(\tau) = 0$ or $F(\tau) > 0$. Suppose that $F(\tau) = 0$, then for any z we have $f(z) = \frac{\frac{\lambda}{2}\phi}{\lambda\phi - \eta} \int_{\tau}^{z-\tau} f_t(y) f_t(z-y)\, dy$ where necessary $\eta < \lambda\phi$. But this implies that $z < 2\tau \Rightarrow f(z) = 0$. Iteratively then one shows that $f = 0$. Therefore it must be the case that $F(\tau) > 0$.

This findings is important as it shows that even arbitrary small clearing rates for large mergers are not enough to prevent concentration.

Theorem 3.1. *For any $p \in \mathcal{R}(\mathcal{Z}^2)$, such that p has a finite support, then $p \in \mathcal{A}_{f_0}(\lambda)$.*

Proof: This result follows from Propositions 4 and 5. For any policy p with finite support, we can find a threshold-based policy with $\phi = 0$ that is more lax than p. The space of active policies is likely not contained in the space of threshold-based rules. We form the following conjecture that we have not yet been able to prove formally

Conjecture 3.1. *For any $p \in \mathcal{R}(\mathcal{Z}^2)$, such that p is integrable, then $p \in \mathcal{A}_{f_0}(\lambda)$?*

3.4.3 Examples

To gain more insight into the precise evolution of a firm size distribution under antitrust regulation we now provide an example admitting closed-formed solutions.

$f_0(t) = \alpha e^{-\alpha t}$ **and** $\phi = 0$**.** There exists a unique stationary solution steady state of the form $f_\infty(z) \sim \beta^*(\tau) e^{-\beta^*(\tau) z}$, where β^* is strictly increasing in τ. A higher threshold leads to larger mean and variance. Remarkably these numbers are independent of α and λ and depending only on τ. See Appendix for details.

3.5 Equilibrium Merger Behavior

We now analyze the equilibrium merger behavior of profit-maximizing firms. These firms must spend resources in the search of merging partners. Antitrust policies affect the intensity at which firms search because it changes the probability that merger deals will go through.

3.5.1 Market structure

A firm with index z produces $y(z)$ units of output and charges a markup $\mu(z)$ over marginal cost. In section 5, we introduce production function that require labor in order to study the evolution of the labor share. The total real output in this economy is

$$Y_t = \int_{z \geq 0} y(z) \, dF_t(z)$$

and total revenues are

$$\Pi_t = \int_{z \geq 0} \mu(z) \, y(z) \, dF_t(z)$$

The aggregate profit ratio in this economy is $\frac{\Pi_t}{Y_t}$. We will focus on a class of production function that permits mergers to create synergies. We also define the output and profit cumulative densitiy functions. Let $F_t^Y(x) := Y_t^{-1} \int_{z \geq 0}^{x} y(z) \, dF_t(z)$ denote the share of total output produced by firm with productivity index up to $x$and $F_t^\Pi(x) := \Pi_t^{-1} \int_{z \geq 0}^{x} \Pi(z) \, dF_t(z)$ their share of profits.

Definition 3.5. A production function y is super-additive if and only if $\forall z_{1,} z_2 \in \mathcal{Z}, y(z_1 + z_2) \geq y(z_1) + y(z_2)$.

Assumption 3: The production function $y : \mathcal{Z} \to \mathbb{R}$ is super-additive

For simplicity, we assume in this section that a firm with productivity index z realizes an output $y(z) = e^z$, where $\mathcal{Z} := [\ln 2, +\infty)$, an interval on which the exponential is super-additive. Therefore the total output in the economy is

$$Y_t = \int_{z \geq 0} e^z f_t(z) \, dz$$

Firms with index z charge a markup over price $\mu(z)$. Profits are $\pi(z) = [\mu(z) - 1] e^z = \bar{\pi}(z) e^z$, where $\mu(z)$ is the markup over marginal costs. Therefore aggregate profits are $\Pi_t = \int_{z \geq 0} \bar{\pi}(z) e^z f_t(z) \, dz$. The revenue-based firm size distribution is therefore $s_t(z) = \frac{1}{M_t} \cdot \bar{\pi}(z) e^z$. Larger firms typically enjoy larger market power than smaller firms. We make the following assumption regarding the markup functions.

Assumption 4: Increasing Markups. The markup function $\mu : \mathcal{Z} \to \mathbb{R}$ is non-decreasing.

This assumption is consistent with several market structures and demand specifications widely used in the literature – monopolistic competition, Kimball preferences, or Cournot competition –. We provide details for several market structures in the Appendix.

3.5.2 Concentration Measures

Concentration Ratios Concentration can be measured in various ways. Concentration indices can be computed as quantiles of the survivor function. For instance, the share of output produced by the firms above the quantile q is, $CR^Y_{q,t} = 1 - F^Y_t(q)$. These quantities can be related to the concentration rations reported in many business statistics report. These empirical concentration ratios are computed in industries containing a finite number of firms. Suppose that there are N firms in a given sector, the share of output $CR^Y_{n,t}$ of the n largest

firms is $CR^Y_{n,t} = 1 - F^Y_t\left(\frac{1-n}{N}\right)$. [14]

Lorenz curve We can also look at the Lorenz curve, defined as $L(q) \equiv \frac{\int^{F^{-1}(q)} zf(z)dz}{m^1}$ for $q \in [0,1]$. The Lorenz curve evaluated at q captures the share of output commanded by the all the firms up to the q^{th} quintile. If output was produced linearly then $L^Y_t(q) = q$. As long as the production function is convex in z the Lorenz curve is convex as well. Similarly we define the profits Lorenz curve L^Π_t.

Herfindahl index The Herfindahl index is usually defined for a finite number of firms as the sum of the squared market shares. A continuous counterpart of the Herfindahl index is $HH_t = \int_{z\geq 0}\left[100 * \frac{\mu(z)y(z)}{\Pi_t}\right]^2 dF_t(z)$.

3.5.3 Equilibrium Search

The analysis has so far taken the search rate to be an exogenous parameter. It is now endogenized by supposing that firms choose their search level at any time, that is, firms choose a process $\lambda_t \in \left[\underline{\lambda}, \overline{\lambda}\right]$ that is progressibely measurable with respect to the filtration generated by the random meeting process. At any point in time the cost of reaching a search level λ_t is $\chi(\lambda_t)$, where χ is a non-negative, increasing and convex function. The superadditivity assumption in productivities garanties that mergers in our model always create value for the two parties involve. Therefore the value function, V_t are increasing. We make the following assumption regarding how the surplus is split.

Assumption 5 (Merger Surplus): *When two firms merge, the target firm receives a fraction $\beta \in (0,1)$ of the surplus created by the merger.*

Therefore when a firm with index z buys a firms with index y the target receives $\beta\left[V_t(z+y) - V_t(y)\right]$. This assumption, standard in the search literature (e.g. Hugonnier, Lester, and Weill 2018), can be given formal microfoundations with a Nash-bargaining set-

[14]For instance in the US, the concentration ratio for the 4,8, 50 largest firms are reported by the US Census at the 2-digit SIC level at least.

ting. There is no strong evidence in favor of any particular value for β. It is convenient to make the following symmetry assumption to simplify some of the algebra. This assumption is not critical to derive any of the results but notice that giving more bargaining power to the acquirer necessarily increases the incentives to search for large firms, who are more likely to fall on the acquiring side. [15]

Assumption 6 (Symmetric Bargaining Power): *The merger gains are split evenly:* $\beta = \frac{1}{2}$.

An application of Bellman's principle of optimality shows that the value function of a firm with productivity z at time is

$$V_t(z) = \max_{\lambda} \mathbb{E}_t \left[\int_t^{\tau} e^{-rx} \left[\pi(z_x) - \chi(\lambda_x) \right] dx + \beta \frac{1}{2} \int_0^{\infty} \left[\left[V_t(z+y) - V_t(z) \right] p(z,y) f_t(y) \, dy \right] \right] \tag{3.20}$$

where $\tau$is an exponential random variable with parameter λ representing the arrival of the first meeting. The value of the firm increases continuously with the flow profits, to which search costs are substracted. The differential Bellman equation associated to the firm maximization problem in the presence of antitrust policy p is therefore

$$\begin{aligned} rV_t(z) \quad = \max_{\lambda} \quad & \pi(z) - \chi(\lambda) + \underbrace{\lambda\beta \int_{y>z_t} \left[V_t(z+y) - V_t(z) \right] p(z,y) f_t(y) \, dy}_{\text{Acquired in a Merger by a larger competitor}} \\ & \underbrace{+ \lambda(1-\beta) \int_{y<z_t} \left[V_t(z+y) - V_t(z) \right] p(z,y) f_t(y) \, dy}_{\text{Acquires a smaller competitor}} \end{aligned} \tag{3.21}$$

This forms make explicit the two kinds of events can make the value jump: the firm can be acquired by a larger competitor, in which case it captures a share β of the surplus created by the merger, or, the firm meets a smaller competitor, it ends up on the acquiring side and gets a share $(1-\beta)$ of the surplus.

[15] By changing the balance between acquirers and targets' incentives to search one affects the level and the convexity of the equilibrium search function.

Symmetric bargaining power can result from a game-theoretical setting where the buyers and sellers make take-it-or-leave-it offers alternatively. The Bellman equation simplifies to

$$rV_t(z) = \max_{\lambda} \pi(z_t) - \chi(\lambda) + \lambda \tfrac{1}{2} \left[\int_0^\infty [V_t(z+y) - V_t(z)] p(z,y) f_t(y) dy \right] \quad (3.22)$$

Because the gains to search are linear in λ but the costs are convex, there is a unique optimal level of search given by the first order condition in the search effort is

$$\chi'(\lambda) = \frac{1}{2} \int_0^\infty [V_t(z+y) - V_t(z)] p(z,y) f_t(y) dy \quad (3.23)$$

This equation is key to understand the dynamics of the firm size distribution. The incentives to search depend on the depth of the potential merging partner, the potential gains, and the antitrust policy rule. The right hand side is non-negative since V_t is nondecreasing. However, suppose p is equal to 0 when the maximum productivity of the merging entities is above a certain threshold τ. In this case, the right hand side would be equal to 0 and there would be no search above τ. Let $\lambda^*(z)$ denote the optimal search level for a firm with productivity z. The enveloppe condition yields the following equation

$$r \frac{\partial V_t(z)}{\partial z} = \pi'(z) - \chi(\lambda^*(z)) + \lambda^*(z) \tfrac{1}{2} \int_0^\infty \left[\frac{\partial V_t(z+y)}{\partial z} - \frac{\partial V_t(z)}{\partial z} \right] p(z,y) f_t(y) dy \quad (3.24)$$

Let $P_t(z) = \int_0^\infty p(z,y) f_t(y) dy$ denote the expected approval rate for a merger involving a firm of type z.

We now turn to threshold-based policies to analyze the behavior of firms in more details.

Steady state equilibrium Let $P(z) = \int_0^\infty p(z,y) f(y) dy$ define the average probability that a merger is accepted given that it involves a firm of type z. We can rewrite the Bellman equation as follows:

$$\left[r + \lambda^*(z) \tfrac{1}{2} P(z) \right] V(z) = \pi(z) - \chi(\lambda^*(z)) + \lambda^*(z) \tfrac{1}{2} \int_0^\infty V(z+y) p(z,y) f(y) dy \quad (3.25)$$

Taking the function f, P, and λ^* as given, the mapping

$$T[V](z) = \left[r + \lambda^*(z) \frac{1}{2} P(z) \right]^{-1} \left[\pi(z) - \chi(\lambda^*(z)) + \lambda^*(z) \frac{1}{2} \int_0^\infty V(z+y) p(z,y) f(y) dy \right]$$

is a contraction mapping since the Blackwell conditions – discounting and monotonicity – are satisfied. Therefore for any f, P, and λ^* there is a unique value function solution to this problem.

3.5.4 Equilibrium under Threshold-based policies

Suppose that the competition authority implements a threshold-based policy rule $p \in \mathcal{R}_{\mathcal{T}}(\mathcal{Z}^2)$, then for $z > \tau$ the firm problem reduces to

$$\begin{aligned} rV_t(z) \quad = \max_{\lambda} \quad & \pi(z) - \chi(\lambda) + \tfrac{1}{2}\lambda\phi \int_{y>\tau} [V_t(z+y) - V_t(z)] f_t(y)\, dy \\ & + \tfrac{1}{2}\lambda \int_0^{\tau} [V_t(z+y) - V_t(z)] f_t(y)\, dy \end{aligned}$$

We now derive equilibrium counterparts to the theorems proven in section 2, when search intensity was exogenous.

Theorem 3.2. *In the equilibrium merger model, a threshold-based antitrust policy p is active if and only if $\phi = 0$.*

To prove this result, we first show that policies that do not block mergers completely above a threshold lead to equilibrium search rates that are bounded away from 0.

Lemma 3.1. *Suppose that $\phi > 0$, then the equilibrium search rate is bounded below $\lambda^*(z) > \underline{\lambda} > 0$*

Proof: We sketch the main steps of the demonstration 1. using the FOC in λ

$$rV_t(z) = \max_{\lambda} \pi(z) - \chi(\lambda) + \frac{1}{2}\lambda \int_0^{\tau} [V_t(z+y) - V_t(z)] f_t(y)\, dy$$

2. for large values of z , $y \ll z$ and we have that

$$rV_t(z) \approx \max_{\lambda} \pi(z) - \chi(\lambda) + \frac{1}{2}\lambda V_t'(z) \int_0^{\tau} y f_t(y)\, dy$$

3. One can now show that the marginal gains to searching merging partner are bounded away from 0. As long as $V_t'(z)$ is bounded away from 0, for instance if profits are not concave.
4. the results from applying the propositions in Section 2.

Conversely, if $\phi = 0$, for z above the threshold, if there is no search then

$$V_t(z) = \frac{\pi(z)}{r+\delta}$$

3.6 Concluding Remarks

We studied the role of antitrust policies in shaping competition, growth and the firm size distribution. When firms are able to merge, the firm size distribution tend to grows exponentially. Active antitrust policies that block all the mergers above a certain threshold are required in order to make the firm distribution converge. We have focused on developing the mathematical structure of the framework and demonstrating that it can easily accommodate many natural extensions. The next step is to take the model to the data. We are currently collecting extensive data on merger deals to document the empirical relationship between antitrust policies and the firm size distribution. Because the speed of convergence is exponential, a regime change from a passive to an active antitrust policy can rapidly increase the concentration in the economy. Another interesting application of the model would be to study the impact of cross-national mergers involving countries with large productivity gaps.

3.A Extensions of the Model

The baseline model can be extended in several directions. We first examine the interaction between M&A activity and innovation. We then explore alternative assumptions regarding the distribution of entrants, business failures and learning from external sources.

3.A.1 Innovation and Organic Growth

Heterogeneous firms models following Klette and Kortum 2004 emphasize the role of innovation in driving the evolution of firms sizes. We extend our model to integrate this channel. In equilibrium, innovation activity in these models result in a constant drift. To understand the impact of this channel in our model we simply add a drift term to the law of motion, which becomes

$$\frac{\partial f_t(x)}{\partial t} = \underbrace{\eta g_t(x)}_{\text{entry}} + \underbrace{\gamma \frac{\partial f_t(x)}{\partial x}}_{\text{organic growth}} - \underbrace{\lambda f_t(x)}_{\text{learning outflow}} + \underbrace{\frac{\lambda}{2} f_t(x) * f_t(x)}_{\text{learning inflow}}$$

If, as usual, we assume that $g_t = f_t$ and set $\eta = \frac{\lambda}{2}$ in order to keep the mass of firms constant, then taking Laplace transforms of the equation yields

$$\begin{aligned} \frac{\partial \hat{f}_t(s)}{\partial t} &= -\tfrac{\lambda}{2}\hat{f}_t(s) + \gamma s \hat{f}_t(s) + \tfrac{\lambda}{2}\hat{f}_t^{\,2}(s) \\ &= -\left(\tfrac{\lambda}{2} - \gamma s\right)\hat{f}_t(s) + \tfrac{\lambda}{2}\hat{f}_t^{\,2}(s) \end{aligned}$$

with solution

$$\hat{f}_t(s) = \frac{\hat{f}_0(s)}{e^{\tilde{\lambda}(s)t}\left[1 - \frac{1}{\tilde{\lambda}(s)}\frac{\lambda}{2}\hat{f}_0(s)\right] + \frac{1}{\tilde{\lambda}(s)}\frac{\lambda}{2}\hat{f}_0(s)} \tag{3.26}$$

where $\tilde{\lambda}(s) = \frac{\lambda}{2} + \gamma s$.

Proposition 19. *In the presence of additional organic growth,*

i) the mean of the distribution grows at a higher rate but

ii) the speed of convergence is unchanged, in particular, we have

$$m_t^1 = \left(m_0^1 + \frac{\gamma}{\lambda/2}\right) e^{\frac{\lambda}{2}t} - \frac{\gamma}{\lambda/2}$$

While organic growth contributes to raising the mean of the distribution, it does not affect the growth rate.

3.A.2 Learning from an External source

Learning, an important mechanism studied in the "idea flow" literature can be integrated smoothly in our model. Suppose that there is an exogenous sourceof ideas with probability distribution h_0. Firm can learn from that source and use this knowledge to improve on their current productivity.

$$\frac{\partial f_t(x)}{\partial t} = -(\lambda+\nu) f_t(x) + \frac{\lambda}{2}\underbrace{f_t * f_t(x)}_{\text{Mergers}} + \underbrace{\nu f_t * h_0(x)}_{\text{Learning from external source}} + \underbrace{\eta f_t(x)}_{\text{Entry}} \tag{3.27}$$

where $h_0()$ is an external source of "ideas" with which current productivity can be combined. Suppose that $\eta = \frac{\lambda}{2}$ such that the mass of firms is constant. We transform this partial differential equation using a Laplace transform instead of the Fourier transform in order to avoid having to deal with complex numbers. We obtain the following ordinary differential equation:

$$\frac{\partial \hat{f}_t(s)}{\partial t} = -(\lambda+\nu-\eta)\hat{f}_t(s) + \frac{\lambda}{2}\hat{f}_t^2(s) + \nu \hat{f}_t(s)\cdot \hat{h}_0(s)$$

whose solution can be written compactly as

$$\hat{f}_t(s) = \frac{\hat{f}_0(s)}{e^{\tilde{\lambda}(s)t}\left[1 - \frac{1}{\tilde{\lambda}(s)}\frac{\lambda}{2}\hat{f}_0(s)\right] + \frac{1}{\tilde{\lambda}(s)}\frac{\lambda}{2}\hat{f}_0(s)}$$

where $\tilde{\lambda}(s) \equiv \lambda + \nu\left[1-\hat{h}_0(s)\right] - \eta = \frac{\lambda}{2} - \nu\left[1-\hat{h}_0(s)\right]$ or in Wild sum

$$\hat{f}_t(s) = \sum_{n\geq 1} e^{-\tilde{\lambda}(s)t}\left[\frac{\lambda}{2}\frac{1}{\tilde{\lambda}(s)}\left(1-e^{-\tilde{\lambda}(s)t}\right)\right]^{n-1}\hat{f}_0(s)^n$$

While the speed of convergence is affected by the characteristics of the source distribution, the main structure of the model is preserved, which allows us to compare the dynamics of this economy with the baseline economy's.

Proposition 20. *The mean of the distribution grows asymptotically at the rate* $e^{\frac{\lambda}{2}t}$*, in particular, we have*

$$m_t^1 = e^{\frac{\lambda}{2}t}\left[m_0^1 + \frac{\nu}{\lambda/2}m_h^1\right] - \frac{\nu}{\lambda/2}m_h^1 \tag{3.28}$$

This results means that, as long as the external source has a fixed distribution, its presence does not substantially alter the long-run growth of this economy. However the presence of the external is important initially, especially if its mean high realtive to the mean productivity in the economy. An alternative and perhaps more interesting experiment is to consider an economy with an external source of knowledge but where the overall meeting rate is the same as in the baseline economy without external source of knowledge. In this case the external source crowds out mergers, firms are randomly meeting merging partner at rate $\lambda' = \lambda - \nu$. We have the following result

Proposition 21. *The mean of the distribution grows asymptotically at the rate* $e^{\left(\frac{\lambda-\nu}{2}\right)t}$*, in particular, we have*

$$m_t^1 = e^{\left(\frac{\lambda-\nu}{2}\right)t}\left[m_0^1 + \frac{\nu}{\frac{\lambda-\nu}{2}}m_h^1\right] - \frac{\nu}{\frac{\lambda-\nu}{2}}m_h^1 \tag{3.29}$$

The asymptotic growth rate is now strictly smaller than in the baseline economy, However in the presence of the external source the economy can enjoy a higher growth rate initially if the mean of the external source's distribution is high enough.

If firms could choose between searching for a merging party and learning from the external source, they would all prefer to search in the external source until the mean of that source gets overcome by the mean productivity of the firm's distribution, at which point all firms would now all be searching for mergers.

3.A.3 Entry

Case 1: Barriers to entry $\eta_t = 0,\ \forall t$**.** Suppose that the barriers to entry are so high that no firm enter, the mass of firms decline at a steady rate because of concentration and the firm-size distribution becomes degenerate. We have a Riccati equation with constant

coefficients for any $s \in \mathbb{R}$. The unique solution is

$$\hat{f}_t(s) = \frac{\hat{f}_0(s)}{e^{\lambda t}\left[1 - \frac{1}{2}\hat{f}_0(s)\right] + \frac{1}{2}\hat{f}_0(s)} \tag{3.30}$$

where $\hat{f}(s,0)$ is the Fourier transform of the initial distribution of productivity $f_0(z)$. To obtain an analytical expression for f, we can expand the expression into a Wild summation:

$$\hat{f}(s,t) = e^{-\lambda t}\hat{f}_0(s)\sum_{n\geq 1}\left[\frac{1}{2}\left(1 - e^{-\lambda t}\right)\hat{f}_0(s)\right]^{n-1}$$

for any $z \in \mathbb{R}$ and then by linearity of the inverse Fourier transform we have:

$$f(z,t) = \sum_{n\geq 1} e^{-\lambda t}\left[\frac{1}{2}\left(1 - e^{-\lambda t}\right)\right]^{n-1} f_0(z)^{*n} \tag{3.31}$$

where $*n$ is the n-fold convolution operator[16]. The measure f is a mixture of convolutions of the initial distribution. The interpretation is as follows. At any time t, a firm with index z can only be the result of a *finite* number n of successive mergers and the number of such meetings since the beginning of time is given by $e^{-\lambda t}\left[\frac{1}{2}\left(1 - e^{-\lambda t}\right)\right]^{n-1} M_0$. Because our model can be solved in analytically, we are able to characterize the speed at which this distribution converges, which is exponential.

Proposition 22. *If there are barriers to entry ($\eta = 0$), then*

i) there is no stationary solution,

ii) the mass shifts towards infinity: $\forall z, \lim_{t\to\infty} F(z,t) = 0$.

iii) the measure converges exponentially at rate λ,

$$\exists C_1, C_2 > 0, \forall z,\ C_1 e^{-\lambda t} \leq f(z,t) \leq C_2 e^{-\lambda t} \tag{3.32}$$

The moments of the distribution can also be derived analytically using the Fourier transforms. The reason for this is that the Fourier transform of a probability distribution is the conjugate of its characteristic function, defined as $\varphi_f(s) \equiv \int e^{-isx} f(x)\,dx$. Thus, the moments of the distribution can be expressed as follows $\mathbb{E}[X^k] = (-i)^k \varphi_f^{(k)}(0) = i^k \hat{f}^{(k)}(0)$. Let $m_t^k := \mathbb{E}[X_t^k]$ denote the k^{th} uncentered moment of the distribution.

[16]By convention $f_0(z)^{*1} = f_0(z)$

Proposition 23. *If there are barriers to entry ($\eta = 0$), then*

i) the mean of the distribution grows exponentially at rate $\frac{\lambda}{2}$, in particular

$$m_t^1 = m_t^1 \cdot e^{\frac{\lambda}{2}t} \cdot \frac{1}{\left[1 - \frac{1}{2}M_0\left(1 - e^{-\lambda t}\right)\right]^2}$$

ii) the variance of the distribution shrinks exponentially at rate $-\frac{\lambda}{2}$, in particular

$$Var\left[Z_t\right] \sim Var\left[Z_0\right] \cdot e^{-\frac{\lambda}{2}t} \cdot \frac{1}{\left[1 - \frac{1}{2}M_0\left(1 - e^{-\lambda t}\right)\right]^2}$$

Case 2: Exogenous entry distribution $g_t(z) = g_0(z)$. Section 2 assumes that entrants are able to access the same distribution of technologies than established firms, a polar assumption is that entering firms draw their productivities from a fixed external distribution g_0. The law of motion is

$$\frac{\partial \hat{f}_t(s)}{\partial t} = -\lambda \hat{f}_t(s) + \frac{\lambda}{2}\hat{f}_t^{\,2}(s) + \frac{\lambda}{2}\hat{g}_0(s)$$

which is a Ricatti equation with constant coefficients. The Fourier transform of any stationary distribution f_∞ must solve the quadratic equation

$$\hat{f}_\infty^2(s) - 2\hat{f}_\infty(s) + \hat{g}_0(s) = 0\,,\ \forall s$$

This equation has two roots but since we have $\left\|\hat{f}(s)\right\| \leq 1$, by property of the Fourier transform, the correct solution is $\hat{f}_\infty(s) = 1 - \sqrt{1 - \hat{g}_0(s)}$. [17]The Fourier transform of the solution is therefore

$$\hat{f}_t(s) = \hat{f}_\infty(s) + \frac{\left[\hat{f}_0(s) - \hat{f}_\infty(s)\right] \cdot \Lambda(s)}{\Lambda(s) + \left[\hat{f}_0(s) - \hat{f}_\infty(s) + \Lambda(s)\right]\left(e^{\frac{\lambda}{2}\Lambda(s)t} - 1\right)} \tag{3.33}$$

where $\Lambda(s) = 2\sqrt{1 - \hat{g}_0(s)}$. Notice that while the solution converges to a stationary solution f_∞, the latter has an infinite mean since $\lim_{s\to 0}\hat{f}'_\infty(s) = \lim_{s\to 0}\frac{1}{2}\hat{g}'(s)\left[1 - \hat{g}(s)\right]^{-\frac{1}{2}} = +\infty$.

[17]The condition that $\hat{g}_0(s) \leq 1, \forall s$ is satisfied for the same reason by property of the Fourier transform. However, complications arise if the Fourier transform evaluted at s is complex-valued. In these case one can use the Laplace transform instead.

Case 3: Entry with lagged Productivity $g_t(z) = f_t(z+\zeta)$. We can embed limits to knowledge transmission in our model by assuming that entering firms do not draw from the current firm size distribution. Instead, entrants draw from a distribution that is shifted downwards by an amount ζ. The PDE in the Laplace transform is

$$\frac{\partial \hat{f}_t(s)}{\partial t} = -\lambda \hat{f}_t(s) + \frac{\lambda}{2}\hat{f}_t^{\,2}(s) + \eta e^{s\zeta}\hat{f}_t(s)$$

with solution

$$\hat{f}(s,t) = \frac{\hat{f}(s,0)}{e^{\tilde{\lambda}(s)t}\left[1 - \frac{1}{\tilde{\lambda}(s)}\frac{\lambda}{2}\hat{f}(s,0)\right] + \frac{1}{\tilde{\lambda}(s)}\frac{\lambda}{2}\hat{f}(s,0)}$$

where $\tilde{\lambda}(s) = \frac{\lambda}{2} - \eta e^{s\zeta}$. The solution for $\eta = \frac{\lambda}{2}$ is

$$\hat{f}_t(s) = \sum_{n\geq 1} e^{-\tilde{\lambda}(s)t}\left[\frac{\lambda}{2}\frac{1}{\tilde{\lambda}(s)}\left(1 - e^{-\tilde{\lambda}(s)t}\right)\right]^{n-1}\left(\hat{f}_0(s)\right)^n$$

Case 4: Entry with Scaled-down productivity $g_t(z) = f_t(\kappa \cdot z)$ with $\kappa > 1$. Instead of drawing from the current productivity distribution, suppose that entrants draw from a scaled-down version of the distribution. If f_t is decreasing then for any z, $f_t(\kappa \cdot z) \leq f_t(z)$. The differential equation is

$$\frac{\partial \hat{f}_t(s)}{\partial t} = -\lambda \hat{f}_t(s) + \frac{\lambda}{2}\hat{f}_t^{\,2}(s) + \frac{\lambda}{2}\frac{1}{\kappa}\hat{f}_t(\frac{s}{\kappa})$$

3.A.4 Business Failures

The model can accomodate business failures. In the baseline model where entrants draw from f_t, exits can be subsumed under η, which must then be interpreted as the net entry rate. However, even in this case, exits affect the value function of the firm, since it acts as an additional discount rate, and therefore changes the equilibrium search intensity. For instance, suppose that firms fail uniformly at rate δ, then the law of motion becomes

$$\frac{\partial f(x,t)}{\partial t} = \underbrace{\eta g_0(x)}_{\text{entry}} - \underbrace{\delta f(x,t)}_{\text{exit}} - \underbrace{\lambda f(x,t)}_{\text{learning outflow}} + \underbrace{\frac{\lambda}{2} f(x,t) * f(x,t)}_{\text{learning inflow}}$$

where $g_0()$ is an exogenous entry distribution. The model can be analyzed in the same fashion.

3.A.5 Heterogenous Diffusion

The assumption that all firms share the same meeting technology can be relaxed. For instance firms who belong to markets that are separated geographically (e.g. different countries) or correspond to different production sectors using different technologies might be less likely to merge. In fact, empirical evidence shows the merge who merge tend to belong to closely related sectors Suppose that they are M types of agent and that meeting rates, κ_{ij}, depend on the types of a firm pair. For instance, one can divide the economy in sectors and use the normalized input-output matrix to inform κ_{it}. The distribution of firms of type i at time t is given by

$$\frac{\partial f_{it}}{\partial t} = \underbrace{(1-\lambda_i) f_{it}}_{\text{no meeting}} + \underbrace{\frac{\lambda_i}{2} f_{it} * \sum_{j=1}^{M} \kappa_{ij} f_{jt}}_{\text{meetings with each agent class}} \quad , \ i \in \{1, ..., M\}$$

where "meeting" matrix $K = [\kappa_{ij}]_{ij}$ must be a stochastic matrix. Using the same technique, the solution can be developped as a mixture of convolution of the initial distributions in each type.

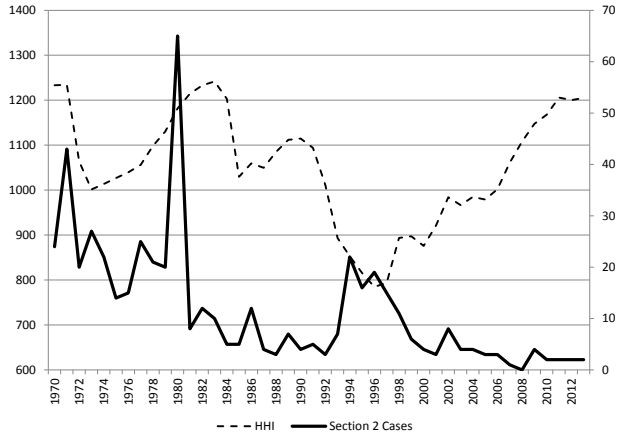
$$f_{i,t} = \sum_{k \in \mathbb{N}^M} a_{i,t}(k) \, f_{1,0}^{*k_1} * ... * f_{M,0}^{*k_M}$$

We leave the detailed study of this extension of the model for future research.

3.B Figures

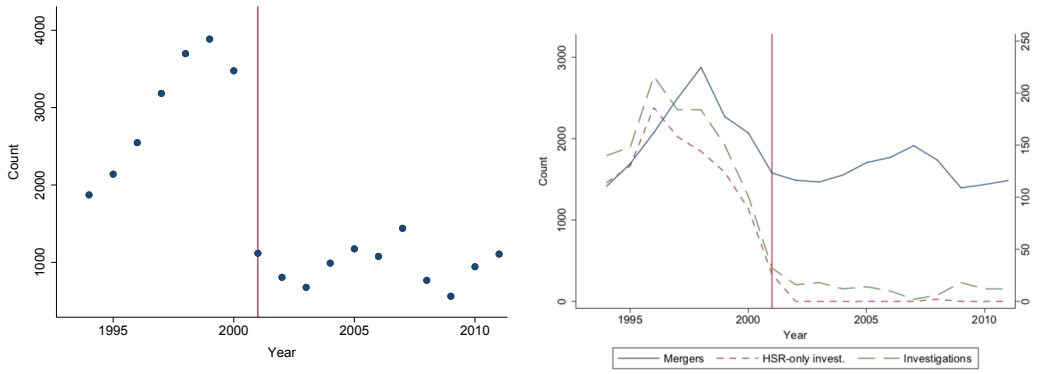
These Figures provide suggestive evidence of 1) the connection between the stance of antitrust policies and concentration; and, 2) the fact that pre-merger notification thresholds act in practice as antitrust policy thresholds, we suggests that they could be used as a reliable proxy.

Figure 3.1: Increasing concentration and decreasing enforcement in the US.



This graph plots the number of cases filed by the Department of Justice under Section 2 of the Sherman Act and the Herfindahl concentration index for all US publicly-traded firms that appear in CRSP and Compustat. Grullon, Larkin, and Michaely 2015

Figure 3.2: US Notification thresholds as de facto policy rules



The graph on the left hand side plots the number of transactions for which pre-merger notifications were filed in the US over time. In 2001, deals valued at less than $50 million were exempted from notification (the red line). Exemption from pre-merger notification acted as a de facto probability of approval of 100%. Source: Wollman (2019)

3.C Mathematical Appendix

This appendix contains the proofs of the propositions and theorems in the main text.

3.C.1 Proof for Baseline Model

Notice that the denominators of the analytical expressions of the Fourier transform solutions are always equal to one when evaluated at $s = 0$, greatly simplifying the calculation of moments. This is because when f is a probability density, $\hat{f}_0(0) = 1$.

Evolution of moments of baseline model. Taking the derivative with respect to s yields

$$\hat{f}_t{}'(s) = \frac{\hat{f}_0{}'(s)\, e^{\frac{\lambda}{2}t}}{\left[e^{\frac{\lambda}{2}t}\left[1 - \hat{f}_0(s)\right] + \hat{f}_0(s)\right]^2}$$

and evaluating at $s = 0$ yields

$$\hat{f}_t{}'(0) = \hat{f}_0'(0)\, e^{\frac{\lambda}{2}t} \tag{3.34}$$

The second derivative, corresponding to the second uncentered moment, is

$$\hat{f}_t{}''(s) = \frac{\hat{f}_0{}''(s)\, e^{\frac{\lambda}{2}t}\left[e^{\frac{\lambda}{2}t}\left[1 - \hat{f}_0(s)\right] + \hat{f}_0(s)\right] - \hat{f}_0{}'(s)\left[1 - e^{\frac{\lambda}{2}t}\right]}{\left[e^{\frac{\lambda}{2}t}\left[1 - \hat{f}_0(s)\right] + \hat{f}_0(s)\right]^3}$$

which evaluated at $s = 0$ yields

$$\hat{f}_t{}''(0) = 2\left(\hat{f}_0'(0)\right)^2 e^{\lambda t} + e^{\frac{\lambda}{2}t}\left[\hat{f}_0''(0) - 2\left(\hat{f}_0'(0)\right)^2\right]$$

Therefore the variance is $\hat{f}_t{}''(0) - \left[\hat{f}_0{}'(s)\right]^2$.

Example 2: Exponential Distribution $f_0(z) = \alpha e^{-\alpha z}, z \geq 0$ We have $\hat{f}_0(s) = \frac{\alpha}{\alpha + is}$. and substituting in equation 3.8 yields

$$\begin{aligned}\hat{f}_t(s) &= \frac{\alpha}{e^{\frac{\lambda}{2}t}[\alpha+is-\alpha]+\alpha} \\ &= \frac{\alpha}{e^{\frac{\lambda}{2}t}is+\alpha} \\ &= \frac{\alpha e^{-\frac{\lambda}{2}t}}{is+\alpha e^{-\frac{\lambda}{2}t}}\end{aligned}$$

which is the Fourier transform of the exponential distribution with parameter $\alpha e^{-\frac{\lambda}{2}t}$.

3.C.2 Proof for Antitrust policy

Exponential distribution and $\phi = 0$

Guessing a solution of the form $f(z) = \beta e^{-\beta z}$ and setting the time derivative to zero, we have

$$\beta e^{-\beta z}\left[\eta - \lambda\left(1 - e^{-\beta\tau}\right) + \lambda\tau\beta\right] = 0$$

However η must balance out the mass lost in the merger so $\eta = \frac{\lambda}{2}F(\tau)$. Therefore the solution β^* is

$$\beta^* = \frac{1}{2\tau}\left(1 - e^{-\beta\tau}\right)$$

Taking derivatives, we have that $\beta^*(\tau)$ is increasing in τ.

Exponential distribution and $\phi > 0$

Guess a solution of the form $f(z) = \Gamma(\alpha)^{-1}\beta^\alpha x^{\alpha-1}e^{-\beta z}$ and set the time derivative to zero

$$\beta e^{-\beta z}\left[\eta - \lambda\left(1 - e^{-\beta\tau}\right) - \lambda\phi e^{-\beta\tau} + \lambda\tau\beta + \frac{\lambda}{2}\phi(z - 2\tau)\beta\right] = 0$$

3.C.3 Proof for Extended Models (Section 5)

Innovation/Organic Growth. The solution to the Fourier differential equation is

$$\hat{f}_t(s) = \frac{\hat{f}_0(s)}{e^{\tilde{\lambda}(s)t}\left[1 - \frac{1}{\tilde{\lambda}(s)}\frac{\lambda}{2}\hat{f}_0(s)\right] + \frac{1}{\tilde{\lambda}(s)}\frac{\lambda}{2}\hat{f}_0(s)} \tag{3.35}$$

where $\tilde{\lambda}(s) = \frac{\lambda}{2} + \gamma s$. The first derivative is

$$\begin{aligned} \hat{f}_t'(s) &= \frac{\hat{f}_0'(s)\left[e^{\tilde{\lambda}(s)t}\left[1-\frac{1}{\tilde{\lambda}(s)}\frac{\lambda}{2}\hat{f}_0(s)\right]+\frac{1}{\tilde{\lambda}(s)}\frac{\lambda}{2}\hat{f}_0(s)\right]}{\left[e^{\tilde{\lambda}(s)t}\left[1-\frac{1}{\tilde{\lambda}(s)}\frac{\lambda}{2}\hat{f}_0(s)\right]+\frac{1}{\tilde{\lambda}(s)}\frac{\lambda}{2}\hat{f}_0(s)\right]^2} \\ &- \frac{\hat{f}_0(s)\left[\gamma i t e^{\tilde{\lambda}(s)t}\left[1-\frac{1}{\tilde{\lambda}(s)}\frac{\lambda}{2}\hat{f}_0(s)\right]+\frac{\lambda}{2}\left(e^{\tilde{\lambda}(s)t}-1\right)\left[\frac{\gamma i}{\tilde{\lambda}(s)^2}\hat{f}_0'(s)-\frac{1}{\tilde{\lambda}(s)}\hat{f}_0'(s)\right]\right]}{\left[e^{\tilde{\lambda}(s)t}\left[1-\frac{1}{\tilde{\lambda}(s)}\frac{\lambda}{2}\hat{f}_0(s)\right]+\frac{1}{\tilde{\lambda}(s)}\frac{\lambda}{2}\hat{f}_0(s)\right]^2} \end{aligned} \tag{3.36}$$

and therefore taking $s \to 0$ and using the fact that $\hat{f}_t(0) = 1$ yields

$$m_t^1 = \left(m_0^1 + \frac{\gamma}{\lambda/2}\right) e^{\frac{\lambda}{2}t} - \frac{\gamma}{\lambda/2}$$

Barriers to entry. The first derivative of the Fourier transform is

$$\hat{f}_t'(s) = \frac{\hat{f}_0'(s)\, e^{\lambda t}}{\left[e^{\lambda t}\left[1-\frac{1}{2}\hat{f}_0(s)\right]+\frac{1}{2}\hat{f}_0(s)\right]^2}$$

and the second derivative is

$$\hat{f}_t''(s) = \frac{\hat{f}_0''(s)\, e^{\lambda t}\left[e^{\lambda t}\left[1-\frac{1}{2}\hat{f}_0(s)\right]+\frac{1}{2}\hat{f}_0(s)\right] - \frac{1}{2}\hat{f}_0'(s)\left[1-e^{\lambda t}\right]}{\left[e^{\lambda t}\left[1-\frac{1}{2}\hat{f}_0(s)\right]+\frac{1}{2}\hat{f}_0(s)\right]^3}$$

Here we have to re-normalize f since $\int f_t = M_t = M_0 e^{-\frac{\lambda}{2}t}$. This also implies that far from declining, GDP grows at exponential rate.

Learning from external source. The derivative of the Fourier transform is

$$\hat{f}_t'(s) = \frac{e^{\tilde{\lambda}(s)t}\left[\hat{f}_0'(s) + \nu\hat{h}_0'(s)\,\hat{f}_0(s)\left[1-\left(\frac{1}{\tilde{\lambda}(s)}-1\right)\hat{f}_0(s)\frac{\lambda/2}{\tilde{\lambda}(s)}\right]\right] + \hat{f}_0(s)^2 + \nu\hat{h}_0'(s)\frac{\lambda/2}{\tilde{\lambda}(s)^2}}{\left[e^{\tilde{\lambda}(s)t}\left[1-\frac{1}{\tilde{\lambda}(s)}\frac{\lambda}{2}\hat{f}_0(s)\right]+\frac{1}{\tilde{\lambda}(s)}\frac{\lambda}{2}\hat{f}_0(s)\right]^2}$$

where $\tilde{\lambda}(s) = \frac{\lambda}{2} - \nu\hat{h}_0(s)$. And therefore

$$\hat{f}_t'(0) = e^{\left(\frac{\lambda}{2}-\nu\right)t}\left[\hat{f}_0'(0) + \frac{\nu}{\lambda/2}\hat{h}_0'(0)\right] + \frac{\nu}{\lambda/2}\hat{h}_0'(0)$$

3.D Alternative Market Structures

In the main text we do not specify the markup function $\mu(z)$ and only require it to be non-decreasing in z. Below we provide several market structures and the associated markup functions.

Monopolistic competition Under monopolistic competition, all firms set the same constant markup $\mu(z) = \mu$, which also correspond to the share of profits in this economy.

Kimball Preferences A popular formulation of heterogenous markups is the Kimball (1995) aggregator. Kimball assumes that a composite good is created using the Kimball aggregator

$$\int \Upsilon(y(z)) f(z) dz = 1$$

where $\Upsilon(1) = 1, \Upsilon' > 0$ and $\Upsilon'' < 0$. This formulation ensures elasticities of substitution between goods that are decreasing in the quantity consumed. Klenow and Willis use the following aggregator function

$$\Upsilon(x) = 1 + (\theta - 1) \exp(1/\epsilon) \epsilon^{\frac{\theta}{\epsilon}-1} \left[\Gamma\left(\frac{\theta}{\epsilon}, \frac{1}{\epsilon}\right) - \Gamma\left(\frac{\theta}{\epsilon}, \frac{x^{\frac{\epsilon}{\theta}}}{\epsilon}\right) \right]$$

which leads to the markup function $\mu(q) = \frac{\sigma}{\sigma - q\frac{\epsilon}{\sigma}}$ where $q = \frac{y}{Y}$ and therefore

$$\mu(z) = \frac{\sigma}{\sigma - \left[\frac{y(z)}{Y}\right]^{\frac{\epsilon}{\sigma}}}$$

An alternative is the specification used by Dotsey and King that gives the markup function $\mu(q) = \frac{1}{1-(1-\varkappa)q}$ where $q = \frac{y}{Y}$, that is,

$$\mu(z) = \frac{\sigma}{\sigma - (1 - \varkappa)\frac{y(z)}{Y}}$$

This markup fonction is convex in z

Cournot Competition Finally, we can partition the economy in sectors, each with a finite number of firms K_s. Within each sector, firms engage in Cournot competition as in Atkeson and Burstein 2008, which leads to the following markups

$$\mu(s_i) = \frac{\eta}{\eta - \left[s_i - (1 - s_i)\frac{\eta}{\rho}\right]}$$

where $s_i = \frac{p_i y_i}{\sum p_k y_k}$ is the market share of firm i in sector s.

CHAPTER 4

A Unified Law of Mortality

How do social and economic conditions experienced early in life shape the evolution of health and of mortality rates over the lifetime? To answer this question, we build a parsimonious model of the evolution of health from birth to death that we estimate using cohort life tables from the Human Mortality Database. A key insight of our approach is that the age profile of mortality rates places constraints on the admissible health processes, which allows us to recover the underlying distribution of health. We use the model to predict the effects of WWII and investigate implications for SES gradients.[1]

[1] joint work with Adriana Lleras-Muney

4.1 Introduction

Circumstances in early childhood have lifelong effects on health and mortality, as documented by vast literatures in medicine, demography, sociology, and economics.[2] There is often a large difference between the observed short and long term effects of different shocks – for example it is commonly found that the effects of interventions fade in the short term but reappear later in life. The extent to which these differences are due to mortality selection, or other dynamics, is not well understood. As Almond, Currie, and Duque 2017 note, in the absence of a quantitative model of the evolution of health and mortality from birth to death it is difficult to predict how shocks will affect population outcomes at various ages, and even more delicate to design optimal investment or compensation policies. This paper proposes and estimates such a model using cohort mortality.

Cohort mortality rates exhibit a remarkably consistent U-shaped pattern: high in infancy and old age, and low, but variable, during reproductive ages. The stability and consistency of this shape across human populations and primates has motivated a search for an underlying "law of mortality", pioneered by Gompertz 1820.[3] Our parsimonious dynamic model tracks the evolution of the health distribution among successive cohorts and provides an exceent fit for mortality age-profiles. In the spirit of classic demographic work (Vaupel, Manton, and Stallard 1979), some individuals are born more frail than others and tend to die young. The health distribution of the survivors then evolves according to a simple law of motion that is formally similar to the stochastic processes used to model corporate defaults (Lando 2004). As in the seminal Grossman 1972 model, we treat health as a stock that deteriorates with age but can increase if (health) resources are invested. But unlike Grossman 1972, resources in our model are stochastic. In addition, individuals can die from accidents unrelated to their health status. In its simplest specification the model is characterized by only five parameters and is able to generate the basic age-profile of mortality. While biological causes

[2]See Almond and Currie 2011 and Almond, Currie, and Duque 2017

[3]In a series of seminal demographic studies, Gompertz 1820, Gompertz 1825, Gompertz 1862, Gompertz 1871 documented the linearity of the logarithm of mortality after age 45.

dictate who survives and who dies in childhood and in old age, "external" causes of death play an important role in explaining the level of mortality during the adolescent years.

The model delivers sharp predictions about how in-utero shocks and socio-economic status shape later health and mortality, helping reconcile conflicting findings in the empirical literature. In the absence of compensatory responses, the model predicts that negative in-utero shocks will increase mortality and lower health at every age. But surprisingly the decline in health among the survivors will exhibit a non-monotonic pattern with age: the effects are large initially, fade by adolescence and slowly start rising with age. Similarly, permanent changes in the level (or the variance) of resources result in changes in health and mortality that vary with age in a surprising manner. These dynamics can seem very different whether the researcher expresses them in levels or in percentage changes.

Our model provides an excellent characterization of the age-profiles of mortality for (selected) cohorts born since the early 19th century. Using the method of simulated moments and cohort life tables from the Human Mortality Database (HMD), the estimation recovers five (or more) parameters from each cohort table and can be used to predict life expectancy and conduct counterfactual simulations. The key implication of the model is that the evolution of health in the population can be inferred from its mortality over the lifetime. To our knowledge there is no other model that can accurately predict population mortality rates from birth to death, while providing a law of motion for health at the individual level.

We close by demonstrating that temporary and permanent changes have very different effects on the age-profile of health and mortality. We illustrate our findings by estimating the effects of WWII, a large but temporary shock. Consistent with previous empirical studies, we find that WWII had long lasting "scarring" effects on mortality among war survivors.

4.2 The Evolution of life-cycle mortality since 1816

We study the evolution of health and mortality among French women born between 1816 and 1947 using data from the HMD which provides population and death counts by age, birth-year and gender collected through vital registration systems (birth and death certificates)

and censuses, from 1816 up to 2015.[4] We focus specifically on French women for convenience and because the French data goes back in time to 1816 and covers a large population.[5] Using the population and death counts, we compute mortality rates by age as the number of deaths divided by the population at that age,[6] and use these to compute survival rates (See Appendix A).[7] Cohort life expectancy rose from 40 for the 1816 cohort and to about 69 for the 1923 cohort.

Figure 4.1 shows the logarithm of mortality rates by age, for selected birth cohorts of women between born between 1860 and 1940 of various european countires (panel a) and for France (panel b).[8] It shows that (the logarithm of) mortality has the shape of a "tick mark": high at birth, low among the young, and high and rising almost linearly with age after middle age. Over time, mortality rates have declined for every age: across cohorts the curves are shifted downwards in almost parallel fashion. These patterns are strikingly similar across countries.

The greatest deviations (in logs or proportional terms) from the tick-mark shape occurs during reproductive ages. There are visible "spikes" corresponding to war years, as can be seen for cohorts born around 1920 who experienced WWII from age 19 to age 25. Even in the absence of wars, for example for the cohorts born in 1860, there is a visible rise in mortality after age 15, which demographers refer to as a "hump" (Preston, Heuveline, and Guillot 2000; Thiele 1871).

---

[4]Despite having the longest and highest quality cohort mortality data, the HMD has some important limitations. Migration is not accounted for. The accuracy of the data falls substantially for years during which the territory changed, which often correspond to wars (1861, 1869, 1914, 1920, 1939, 1943, 1945, 1946). The availability and quality of the data for old ages being limited, imputations are made for all ages above 90.

[5]Studies of health and mortality investigate men and women separately. An analysis of gender differences is beyond the scope of this paper.

[6]Technically we are computing probability, rather than the rate, of dying at a given age.

[7]We have no information on the distribution of births and deaths *within* a year and therefore make no adjustments for the fact that the deaths in the first year do not correspond to individuals born in the first year. The HMD reports probabilities (q_x) that make adjustments based on a series of standard assumptions in epidemiology. Our naively computed probabilities are very similar to the ones HMD computes (See Appendix Figure 4.6).

[8]Appendix Figure 4.7 shows all cohorts between 1860 and 1940

Lastly, the most recent cohorts exhibit no visible humps or spikes. The cohort born in 1940 almost looks like a tick-shape, with an adoloscent hump barely visible. This observation motivates our basic model which seeks to describe "natural" mortality in the absence of "external" causes that do not depend on external factors like war, or choices such as whether to have children.

4.3 A parsimonious model of health and mortality

Aging and "natural mortality"

Individuals are born with an initial health level H_0. This initial health endowment differs across individuals in the population and has an unknown distribution. Every period the environment provides resources I to all individuals, which increase H. In addition, individuals in the environment are more or less lucky, and experience an idiosyncratic shock ε_t to their resources. For example in a stationary environment I characterizes the amount of food that a given country produces, but a given person might receive less if for instance rain was unusually low in their location. The variance of ε_t captures how unequal the distribution of resources within the population is. These idiosyncratic shocks are assumed to be i.i.d. every period. Finally the health stock is subject to depreciation every period $d(t)$, which is increasing in t ($d'(t) > 0$) : every period there is a "user cost", reflecting cumulative cell death and organ damage. Together these forces determine the evolution of the health stock.

People die when their stock of health dips below a threshold $\underline{H}$, which is fixed throughout the lifetime and identical for all individuals. Formally let $D_t = \mathbb{I}(H_t \leq \underline{H}, D_{t-1} = 0)$ denote the random variable equal to one if the individual dies in period t. The population's health and mortality is characterized by the following dynamic system:

$$\begin{cases} H_t = H_{t-1} - d(t) + I + \varepsilon_t & \text{if } D_{t-1} = 0 \\ D_t = \mathbb{I}(H_t < \underline{H}, D_{t-1} = 0), \\ D_0 = 0 \end{cases}$$

with $I \in \mathbb{R}$. Note that if $D_t = 1$ then H_t is undefined–we do not observe the health of individuals if they die. Given this model, the mortality rate at time t is $MR_t = P(D_t = 1|D_{t-s} = 0 \, \forall s < t)$.

To make this model tractable, we make the following parametric assumptions. H_0 follows a normal distribution $N(\mu_H, \sigma_H^2)$, consistent with the empirical distribution of birth weights and other traits measured at birth (Wilcox and T Russell 1983). Shocks to resources every period also follow a normal distribution $\varepsilon_t \sim N(0, \sigma_\varepsilon^2)$.[9] Finally the rate of depreciation is increasing over time: $d(t) = \delta t^\alpha$ with $\delta \in (0, \infty), \alpha \in (0, \infty)$.[10]

Health is the latent *unobserved* construct that determines observed mortality. Appendix Figure 4.8 illustrates for the first two periods the dynamic relationship between population health and mortality rates implied by this model. The initial distribution is normal. In the first period it moves right (if I is positive and larger than the aging term) and gets wider (because of ε_t). Then the individuals to the left of the threshold, die. These individuals were either born frail or had large negative shocks. The infant mortality rate (the fraction of individuals that die in the first period) is given by the area under the curve below the threshold. In the second period this truncated distribution moves right again (if I is large relative to $d(1)$). And the population receives a new shock, generating mortality again among those with suficiently large negative shocks. [11]

The stochastic term ε_t, plays an essential role in the model ensuring mortality rates are positive at every age. In its absence there would be no deaths in period 2 – nor in any subsequent period, until the depreciation term becomes large enough to push the leftmost part of the distribution below the threshold. An implication is that the distribution of health at any age (and therefore the mortality rate) is a function of the entire history of shocks and investments, as can be seen from the definition of MR_t, which conditions on

[9] In simulations alternative assumptions (e.g. log normal errors) resulted in counterfactual mortality rates.

[10] While we do not impose it, estimation typically finds $\alpha > 1$ and the depreciation is therefore concave, as hypothesized by Grossman.

[11] We assume aging starts at birth, consistent with evidence that aging markers are different and evolving among children (Wong et al. 2010).

survival in every previous period. A second key feature of the model is the accelerating aging component, which eventually moves the distribution closer and closer to the threshold, ensuring the death of the entire population. This follows Grossman 1972, and is consistent with biological models of senescence (Armitage and Doll 1954 or Pompei and Wilson 2002).[12]

Despite the model's conceptual simplicity, the mortality rate at a given age cannot be expressed in closed-form.[13] In fact, our model is similar to a class of models for corporate' default probability and securities pricing (e.g. Lando 2004). This literature established that, except for the particular case of a constant or linear drift, these models do not admit closed-form solutions. The model must be estimated using numerical methods, which are very sensitive to initial conditions and model assumptions.

Two out of the seven parameters of the model are not identified. To see this, note that the expression for mortality in the first period (shown in Appendix Figure 4.8) is the standard Probit model. Therefore the threshold $\underline{H}$ and the standard deviation of the initial distribution σ_H are not identified: we can subtract $\underline{H}$ and divide by σ_H on both sides of the expression determining the probability of dying, and leave the mortality rate unchanged. Without loss of generality we therefore set $\underline{H} = 0$ and $\sigma_H = 1$.[14]

The model characterizes the biological evolution of health and mortality of a cohort using 5 (rescaled) parameters: one for the mean initial health (μ_0), two govern the aging process (δ, α), and two characterize the effects of environment, in the form of average investments (I) and the variance of these investments or shocks (σ_ε^2). μ_0 is the distance from the threshold of the initial distribution in standard deviations of the initial distribution. All other parameters are also expressed in "standard deviation" units, except for α which is "scale free"– it does not depend on the initial distribution. Proposition 1 in Appendix B proves that these normalized parameters are identified.

---

[12]See Gavrilov and Gavrilova 1991 and Weibull 1951 for attempts at biological microfoundation drawing on reliability theory from engineering.

[13]This would also be true for the continuous-time analogue of our model – a Brownian motion with a nonlinear drift, where death occurs at the first time this diffusion process hits a threshold set at zero.

[14]More precisely we need to normalize 2 out of three parameters. We find it more intuitive to normalize the threshold rather than the initial mean, but this choice is arbitrary.

Health and mortality over the lifetime

We now describe some basic features of the model. Let $\hat{H}_t \equiv \mathbb{E}[H_t \mid H_t > 0]$ denote the average health in the living population with age t and $\sigma_{H_t} \equiv Var[H_t \mid H_t > 0]$ the variance of health among the living.

Proposition 24. *1. Everyone dies with probability 1:* $\lim_{t\to\infty} Pr(H_t = 0) = 1$.

2. For sufficiently high I (relative to σ_ε^2 *and* σ_H^2*) mortality rates declines (up to some age* t_1*) and then increases with age. (If I is sufficiently negative, mortality rates increase from birth.)*

3. For sufficiently high I (relative to δ*), the average health of the living initially increases and eventually decreases with age.*

4. For sufficiently high I (relative to δ *and* σ*), the variance of health among the living initially increases and eventually falls with age.*

Figure 4.2 illustrates the evolution of the distribution of the health stock as the cohorts ages. At birth health is normally distributed but at age 1, the distribution gets truncated at the threshold. The health distribution then shifts right and spreads out until about age 40 (in this example), at which point it starts moving left before concentrating at the threshold. At any given age after infancy and before old age, the distribution of health is roughly gaussian despite repeated truncations, because it is approximately equal to a sum of normal distributions. This is consistent with the observation that health measures like heights, which grow until adulthood, are close to normally distributed.[15]

The model reproduces the age-profile of mortality well (Figure 4.2b). Log-mortality starts high and plummets to low levels by adolescence, remains low and variable until midlife, and starts rising almost linearly with age. Individuals born with low health endowments account for most of the high infant mortality rate, except for a few unlucky babies initially healthy but hit with large negative shocks. In childhood, mortality rates depend mostly on the

[15]For example Limpert, Stahel, and Abbt 2001 show that either a normal or a log normal distribution fits female heights well.

the variance of the shock, and on the size of the mean investment level which pulls the distribution away from the threshold. Eventually the accelerating depreciation overcomes investment and an increasing number of individuals start falling below the threshold.

Health moves in the opposite direction of mortality (Figure 4.2b). [16]Average population health peaks in mid-life, consistent with the evolution of self-reported health by age for cohorts in the US and in the UK.[17] Because of selective mortality, the variance of health increases and then falls, as reported in Case and Angus S Deaton 2005. Within-cohort health inequality behaves like consumption inequality, which also increases with age – because shocks to consumption are not perfectly correlated across individuals (Angus S. Deaton and C. Paxson 1994, Angus S. Deaton and C. H. Paxson 1997). But it falls late in life because of mortality.

Accidental deaths

Many deaths, like accidents or homicides, strike individuals regardless of their health status. These "extrinsic" causes of death can be integrated in the model by simply adding an i.i.d. "accident shock", ν_t , that is independent of the stock of health H_t (just like corporate finance models complement the firms' assets evolution with a "jump to default" component). Then a constant fraction $\kappa \in [0, 1]$ of the population is killed by an accident every period.[18]

Adding accidents increases mortality rates at all ages (See Appendix Figure 4.9), and especially around reproductive ages, when biological causes of death are dampened. Thus accidents become the dominant cause of death (in percentage terms), consistent with em-

[16]see Angus S. Deaton and C. H. Paxson 1998, Case and Angus S Deaton 2005 for the US and (Contoyannis, Jones, and Rice 2004) for the UK.

[17]The model generates morbidity rates that are U-shaped function of age (where morbidity is defined as the fraction of individuals with health below a health threshold but above the death threshold). Contemporary data show that hospitalization days (a proxy for morbidity) are indeed U-shaped (see table P-10 in Disease Control, Prevention, et al. 2014 https://ftp.cdc.gov/pub/Health_Statistics/NCHS/NHIS/SHS/2014_SHS_Table_P-10.pdf).

[18]Intuitively this random accident rate places a floor in the level of mortality that is constant across ages. If all health-related deaths were eliminated, this accident rate would uniquely determine the longevity of the population ($1/\kappa$).

pirical evidence that external causes of death (unintentional injury, suicide and homicide) account for a larger share of mortality, at least today Remund, Camarda, and Riffe 2018.

Relationship to previous literature

Demographers have sought a unified model of mortality at least since the early 19th century (Gompertz 1825). However his model, like much of the following literature (e.g. T. Li and J. Anderson 2013) only account for mortality *after a certain age*, typically after age 30-40. Thus these models are not suited to study how early conditions impact mortality later in life.

There are a few exceptions. An early model proposed by Heligman and Pollard 1980 uses 8 parameters to describe the probability of dying at a given age for all ages. More recently Sharrow and J. J. Anderson 2016 propose a 6 parameter model, which fits the period (not cohort) tables well. Palloni and Beltrán-Sánchez 2016 propose a model with few parameters that tracks "Barker frailty". Our model differs in one fundamental aspect from these models. Like Grossman, we model individual's stock of health and its evolution, rather than directly modeling the mortality rate of a population. This approach allows for an easy characterization of how factors at a given age affect health and mortality at later ages because one can model inputs into health directly and trace their effects.

Our model differs from Grossman's in several important dimensions. First, health evolves *stochastically* because of the random shocks. Thus the age at death is naturally determined in a stochastic manner, without assuming a deterministic horizon. Second our model can explicitly account for initial health conditions and mortality when analyzing the effects of shocks on the health of the living. Our model is otherwise a much simpler version of the original Grossman model, or more recent versions of the Grossman model by Dalgaard and Strulik 2014 or Galama and Van Kippersluis 2018. We do not model utility or how to think of optimal health inputs: the absence of cohort data on incomes, health inputs, and their prices over time limits our ability to empirically estimate a richer model.

No paper in economics that we know of has made use of the age-profile of mortality over

the entire lifetime to make inferences about the evolution of health. Our contribution is simply to note that cohort mortality life tables can be used to identify a basic parametric model of the evolution of health and mortality, upon which more complex but realistic models can be built.

4.4 The effects of permanent changes in lifetime conditions on health and mortality

We now investigate the implications of the permanent differences in lifetime conditions across populations.

Proposition 25. *Changes in any parameter affect the entire path of mortality:*[19]

1. Increasing the investment I or the average health at birth μ_H unambiguously decreases mortality at all ages.[20]

2. *Increasing any of the aging parameters, δ or α, unambiguously increases mortality at all ages.*

3. *An increase in σ_H^2 can increase or decrease the mortality rate. An increase in σ_H^2 increases the mortality rate at young ages if $\delta t^\alpha \leq I$. Ultimately, an increase in σ_H^2 generates selection and reduces mortality among the oldest.*

4. Investment and health at birth are complements: $\frac{\partial^2 MR_t}{\partial I \partial \mu_H} \leq 0$.

We illustrate some of these results graphically for a specific set of parameters, chosen to describe the 1816 cohort. We compare two cohorts, that are identical in all parameters except for one, which we change by 50%.[21] We then plot either the gap or the percentage change in mortality across the two populations using the baseline population as a reference.

[19] *These statements hold in a model where the accident rate is non-zero throughout the lifetime. See Appendix B for proofs.*

[20] *Changing the threshold also affects mortality rates negatively throughout the lifetime.*

[21] Appendix Figure 4.10 and 4.11 shows how changing each parameter affects the mortality and health curves instead of showing the gaps.

In our simulations we use a finite population of 500,000 individuals. We display the curves only up to age 90 to avoid sampling error, which creates large fluctuations in the mortality rates at very old ages.

In-utero shocks

Negative in-utero shocks such as wars, famines and stress, are equivalent to lowering the mean levels of initial health in a population. Figure 4.3a shows the effects on mortality and on the health of the survivors. Lowering initial health results in markedly higher infant and adult mortality. The effects decline monotonically with age in percent (log) terms, but they are u-shaped in levels (panel a).

A lower initial stock also lowers health among the survivors at all ages, in both levels and percentages (panel b), but the pattern is u-shaped. There is a large health decline initially, almost no impact for many years, and increasingly larger impacts as individuals age. Thus the models predicts exactly what Almond, Currie, and Duque 2017 observe across studies: there is a "fade out" in middle age and effects re-appear later in life, and increase with age after middle age. These results suggest that it will always be difficult to estimate any effect of in-utero shocks in middle age, or to identify individuals that have been affected by shocks during reproducive ages for intervention purposes.

These predictions stand in stark contrast with the Grossman model's original predictions – in which early shocks have a decreasing effect on health with age – but in line with what empirical studies have reported (Almond and Currie 2011). This occurs because in our model the depreciation is not a function of the level of the stock H.

Socio-economic conditions and mortality

A large literature documents large and persistent differences in health and mortality – often called "gradients" – across individuals with different levels of socio-economic status such as educational and income level, or occupation and race (Cutler et al. 2012).

Suppose that higher SES leads to greater I throughout life. Figure 4.3b shows mortality

is higher at all ages for those with lower I. In levels the gap are large at birth, decrease to almost 0 in middle ages, and increase after. In logs the effects are hump-shaped, increasing until middle ages and declining thereafter. This results in log-mortality curves that start converging after some point.[22] This is consistent with Chetty, Stepner, et al. 2016, who show that in the US today those with high incomes at age 40 (a measure of permanent income) have lower subsequent mortality relative to those with lower incomes, with log mortality curves that get closer in old age. If we view education as a strong correlate of lifetime resources, this prediction is also consistent with the literature that has documented that the effects of education on mortality fall with age in percentage terms (Hummer and Lariscy 2011), but increase with age in levels (Schiman, Kaestner, and Lo Sasso 2017).

Lower investments also lower the average health at all ages. But the effect first increases with age, and then starts declining once mortality starts rising, in levels and percentage terms. These predictions are consistent with evidence in Case, Lubotsky, and C. Paxson 2002, Currie and Stabile 2003 and House, Lantz, and Herd 2005, who show that the gaps in self-reported health status and morbidity between those born in poor and rich families grow with age, but decline after 65. Schiman, Kaestner, and Lo Sasso 2017 also show similar evidence that education gradients in self-reported health grow between ages 30 and 65 and then appear to fall.

Alternatively individuals with higher education and incomes might have lower rates of aging δ, due to more frequent physical exercise, lower exposure to pollution or lower stress.[23] Appendix Figure 4.12 shows the effects of increasing the depreciation rate. This results in higher levels of mortality all throughout life, but the effects stay imperceptible for many years, then rise rapidly with age, before petering out in old age. For health, the effects are also small at first, then they rise and fall.[24] Interestingly, if we compare the effects of a rise in depreciation and the effects of a decrease in investment, the patterns are very similar after

[22]See panel b of Appendix Figure 4.10.

[23]Recent work (Z. Liu et al. 2019) shows that education and race are associated with lower methylation rates (effective aging rates at the DNA level).

[24]Changing α has similar effects so we do not show them here.

age 40. Thus with health and mortality data from adults *only,* it will be impossible to infer whether SES is affecting annual resources I or the annual depreciation rate δ.

Adolescent hump

We now consider permanent changes occuring later in life. We focus on explaining those associated with adolescence, a period disruptive in many dimensions. Hormonal levels change abruptly, with large biological and behavioral effects. The begining of adulthood in many societies is also marked by marriage, entry in the workforce and new living arrangements.[25] On the mortality curves, these disruptions manifest themselves in an "adolescent hump".

Figure 4.4a simulates the impact of a permanent change in I, δ, σ, κ or $\underline{H}$ or occurring at age 12.[26] These changes all increase mortality, each in a different way. Decreasing I or increasing δ raises mortality for all subsequent ages. But the effect of higher depreciation is barely noticeable at first, so it cannot fit the adolescent hump. Increasing κ generates a floor between ages 12 and 40, and then the profile converges to counterfactual. Increasing σ also generates something close to a floor but the affected cohort's mortality exhibits a cross-over: it has lower mortality than the original cohort after some age. Finally increasing $\underline{H}$ results in a peak and then the profile slowly converges to counterfactual. This peak does not resemble the adolescent hump is but is similar to what is observed during wars (see below). Qualitatively, increases in the variance or in the accident rate appear to best rationalize the adolescent hump though changes in I or in the threshold could also rationalize it.

We estimate these four specifications for the 1816 cohort. We assume adolescence starts at age 16 based on historical estimates from La Rochebrochard 2000 who reports that the onset of menarche occured around 15.8 in 1816. We find that modelling the hump as an accident results in the highest fit in many dimensions (Appendix Table 4.1). For example we predict a life expectancy of 38.28 whereas the actual life expectancy is 38.25. Figure 4.5a illustrates the results and shows that adding the adolescent hump significantly improves the

[25]For a review see Dahl et al. 2018.

[26]Appendix Figure 4.13 shows the effects on health.

fit. The adolescent accident rate is estimated about 9 per thousand, lowering female life expectancy by around 7.5 years for the 1816 cohort.

This basic 6-parameter model provides an excellent the 1860 cohort, and for chimpanzees (Appendix Tables 4.2 and 4.3, and Appendix Figures 4.14 and 4.15).[27] It may seem surprising that reproductive age mortality is best captured as an increase in the accident rate, unrelated to health status. Historically, maternal mortality accounted for most reproductive age deaths. Loudon 2000 shows that poor hygiene and obstetric practices were mostly responsible for infections (sepsis) and hemorrhage–the main causes of maternal mortality. These practices were widespread, so maternal mortality killed both rich/healthy and poor/unhealthy women. Consistent with our findings, contemporary data shows that mortality rate from external causes of death are well approximated by a step function.[28]

We comment briefly on the other estimated parameters. Initial mean health is about 0.86 standard deviations away from the death threshold. Absent any shocks or investment in the first period, infant mortality would have been roughly 15% (instead of 17%). The annual investment is estimate at 0.4, so in the absence of shocks, the health stock would double by age 3. The variance of resources is about 1, roughly equal to the variance of the initial stock of health. The annual depreciation rate δ is equal to 0.0006. But the aging rate is increasing over time exponentially: α is around 1.8.

4.5 Modelling the effects of temporary shocks on health and mortality

We now use our model to understand how temporary shocks affect the profile of mortality.[29] We arbitrarily start by simulating the effect of temporary changes in parameters starting at

[27] It also also robust to several alternative specifications, see Appendix Table **??**.

[28] This is shown for the US in Appendix Figure 4.16.

[29] Appendix Figure 4.17 shows the effects on average health.

age 20 and lasting 10 years.[30]

Each type of shock leaves a unique imprint on the mortality profile of the affected cohort (Figure 4.4b). Temporary decreases in I generate spikes in mortality, similar to those observed for wars. When investment falls, mortality rates start to rise. They peak in the last year of the shock and fall back thereafter. But mortality rates remain elevated throughout the life (relative to the counterfactual of no shock), thus generating "scarring." Increases in the variance of resources also increases mortality during the shock period. But after the shock ends, mortality falls below counterfactual levels because many individuals have larger positive shocks. By contrast, temporary increases in the accident rate immediately increase mortality, but have no permanent effects: mortality goes back to its initial path immediately after the shock ends. Temporary changes in the depreciation have qualitatively the same effect as temporary changes in I as noted earlier.

Finally, an increase in the threshold generates what is known in demography as "harvesting". It results in very high mortality in the first year of the shock. But mortality starts dropping before the shock ends. Once the shock ends, it dips below counterfactual mortality, rises back up and converges to its counterfactual level. This is because all frail individuals are killed when the threshold first increases. And when the threshold is restored to its original (lower) level mortality falls substantially because there are very few individuals close to the threshold. This pattern fit the effects of extreme weather or pollution events, which appear to displace deaths in short term.[31]

The effects of WWII

We use these results to estimate the effect of WWII for the 1921 cohort, who turned 18 when WWII started in 1939. WWII is the longest conflict in our sample lasting six years

[30]For this simulation we are ignoring the adolescent hump for clarity. See Appendix 4.18 for results of shocks at earlier ages.

[31]For both pollution and weather there is evidence of both short term displacement and longer term effects on mortality. For instance see Schwartz 2000 or Zeger, Dominici, and Samet 1999for the effects of pollution. For the effects of very hot or very cold weather see recent articles by Deschenes and Moretti 2009 and Deschênes and Greenstone 2011.

– this should make it easier to distinguish among different type of shocks.[32] We estimate the structural parameters, explicitly allowing for a shock lasting six years, and varying the type of shock. Perhaps not surprisingly given our simulations, we find that the war is best characterized as a decline in health resources I for women.[33]

Figure 4.5 shows the fit for this model. The parameter estimates show I falling from a lifetime value of 0.29 to a value of -0.11 during the war years. This is consistent with evidence showing that GDP, food supplies, and sanitary conditions declined substantially during the war.[34] Infant mortality rates, which are very sensitive to these inputs, rose substantially during this period.[35]

We estimate that the war lowered life expectancy by approximately 5 years for the 1921 cohort. This of course includes a large number of deaths that occured during the war. Conditional on surviving to 1945, life expectancy is 5 years lower than it would have been in the absence of the war, consistent with the predictions for health (Appendix Figure 4.17). These large scarring effects are underestimated because compensatory responses (like the 1948 Marshall plan) likely lessened the effects of the war. Recent work finds results consistent with our findings. Kesternich et al. 2014 study the effects of the WWII across 13 European countries. They find that individuals more exposed to the war experienced worse economic and health outcomes later in life than other survivors who were less exposed. Havari and

[32]It was also very intense. WWII is estimated to have killed around 600,000 individuals in France, about 1.4% of the 1939 population.

[33]Appendix Table 4.4 shows the results. We evaluate the models in terms of how they fit the overall profile of survival and whether they fit the shock itself. Surprisingly, all models provide almost equally good fits of the survival rate (or the log of q) and the predicted life expectancy. However the fit during the war is clearly better matched by the model that assumes WWII was equivalent to a decline in I. Although all models underestimate the mortality rate and the number of deaths during the war, the I-shock makes the smallest mistakes. It underestimates the number of deaths by 21%. By comparison a pure accident model underestimates the number of deaths by 36%, a variance model by 45% and a change in the threshold by 32%.

[34]GDP declined substantially during the war and 20 to 55% of GDP was appropriated by Germans every year of the occupation (Occhino, Oosterlinck, and White 2006). Food was scarce and food rationing began in 1940. Sanitary conditions deteriorated. For example diphteria rates among school-aged children rose from 32.3 per 100,000 (in 1940) to 110 in 1943 (Stuart 1945).

[35]In the HMD infat mortality was 0.063 in 1938. It rose to a high of 0.085 in 1940, the worse year of the war.

Peracchi 2017 and Schiman, Kaestner, and Lo Sasso 2017 report similar findings for WWII. [36]

There are a few limitations to these results. Population and death counts are subject to measurement error during wars because of changes in territory and migration.[37] We assume that the shock started in 1939 and ended in 1945 but rationing continued until 1949 so the decline in resources likely lasted longer. And not all war years were equally difficult. Based on overall mortality rates, 1944 was the worst year. Finally we choose the most parsimonious model to fit the data using a single parameter – but WWII could have affected many parameters.

4.6 Conclusion

All humans are mortal, except in macroeconomic models where there are often infinitely-lived. This assumption is made for tractability but sidestep a lot of lifecycle patterns. This paper proposes a simple model of the evolution of health and mortality over the life course that we hope could be serve as a basis to study the macroeconomic implications of lifecycle dynamics. The basic model has only six parameters and can be easily estimated by using observed cohort mortality rates to infer the underlying evolution of health. It provide a very good approximation of the mortality profile of cohorts born 1816 to 1940, including the adolescent hump and can be used for instance to understand the effect of in utero shocks, SES, and other temporary and permanent shocks occurring at different points in the lifetime. This model can rationalize many patterns in the literature including scarring, harvesting and dynamic treatment effects that fade, increase or do bith with age.

The parametric restrictions of the model allows cohort and period effects to be be separately identified. However, because of these strong assumptions, the paper has limitations.

[36] Other wars also appear to have caused scarring. For instance Costa 2012 documents that surviving soldiers in WWII have higher morbidity and mortality later in life. Other studies have documented scaring effects of war in utero (Almond and Currie 2011). More recentlyLee 2017 presents evidence of substantial health effects in adulthood among cohorts exposed in utero to the Korean War.

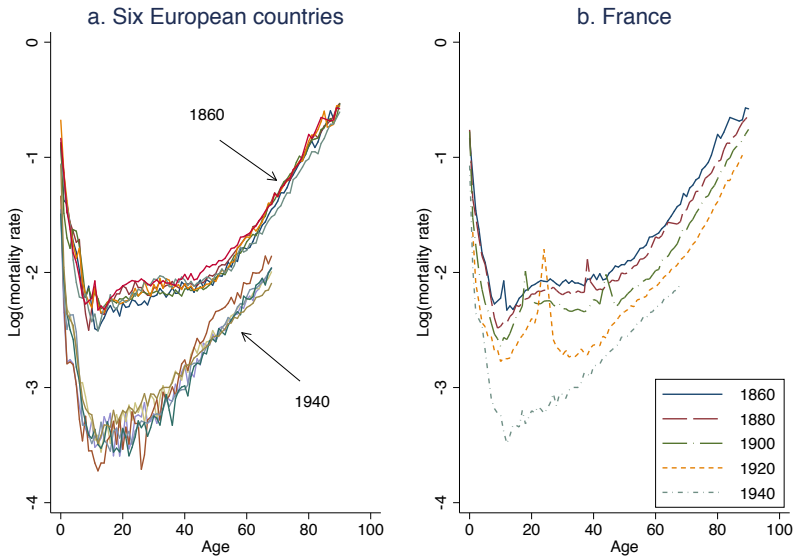
[37] See appendix notes for how we treat the data during these years.

Alternatives assumptions for the distribution of annual shocks or the simple law of motion should be further investigated. The model can also be extended to incorporate behavioral responses and optimization. For example one could make population or individual resources a function of health. With data on inputs, prices and budgets this would be worth modeling and estimating.[38] Better data on health spending by age would make it worth relaxing our assumption that the environment is stationary and it exogenously provides resources, a reasonable assumption for primates or populations with limited technologies, but not for contemporary human populations. The model can be used to investigate many interesting questions that we have not considered such as gender or cross-country differences. One could also use it to study correlations in health across generations in an over-lapping generations where the initial health of a new generation is a function of the average health of mothers during reproductive ages. We leave these to future research.

[38]In preliminary work we found for example that optimal health expenditures are U-shaped over the lifetime.

Figures

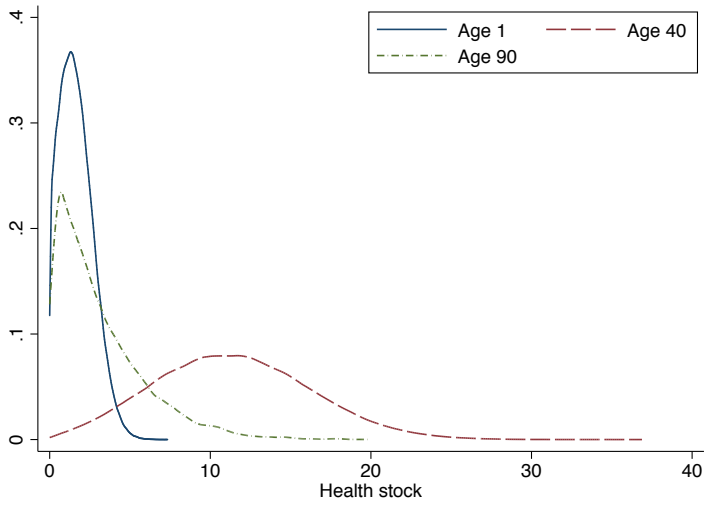
Figure 4.1: The evolution of mortality rates for female cohorts 1860-1940



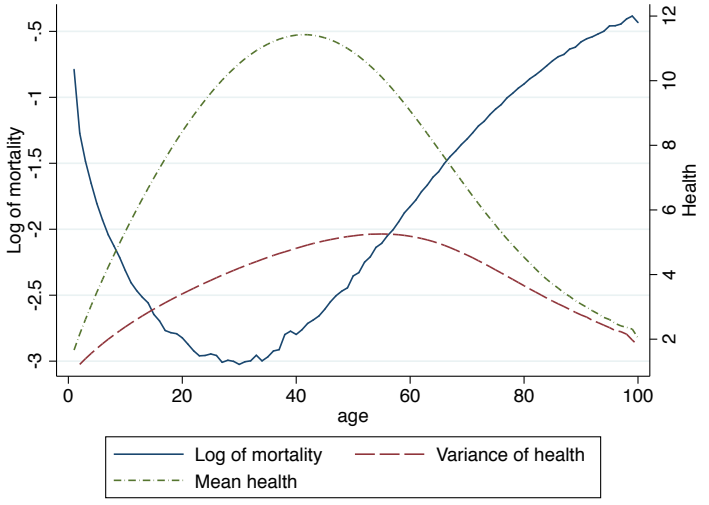
Note: Figures plot the log (base 10) of the mortality rate by age for a given birth cohort using the the Human Mortality Data. Panel a shows data for women in six European countries with data for both 1860 and 1940 (Belgium, Denmark, the Netherlands, Sweden, France, and Norway). Panel b shows several cohorts of French women. We do not show the 1816 data because only one other European country (Sweden) has data for the 1816 cohort.

Figure 4.2: Model behavior

(a) The evolution of the health distribution over the lifetime



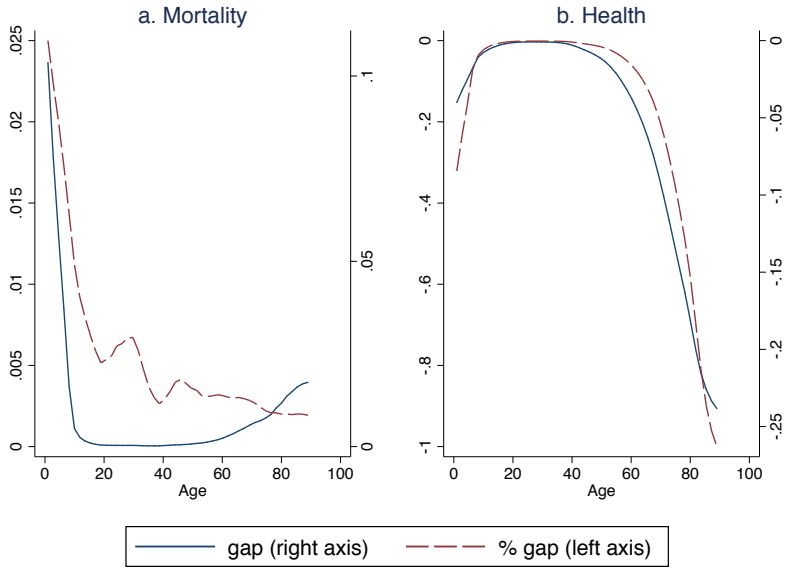
(b) Age profile of population health and mortality



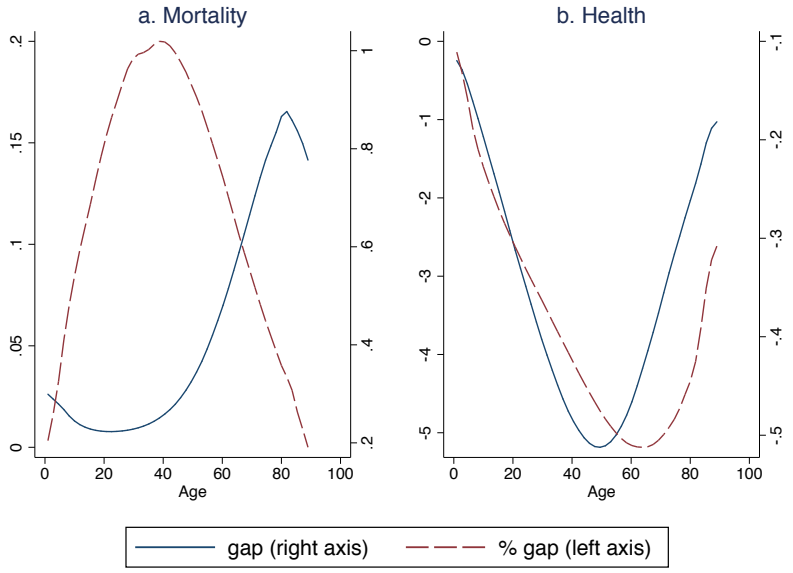
Note: Simulated data for a population of 500,000 individuals. For this simulation we use the following parameters: I=0.3575753, δ=0.0004789, σ=0.8353752, α=1.7883, μ_0=0.925079.

Figure 4.3: Selected Comparative statitics

(a) Effects of decreasing initial health μ_0 by 50%



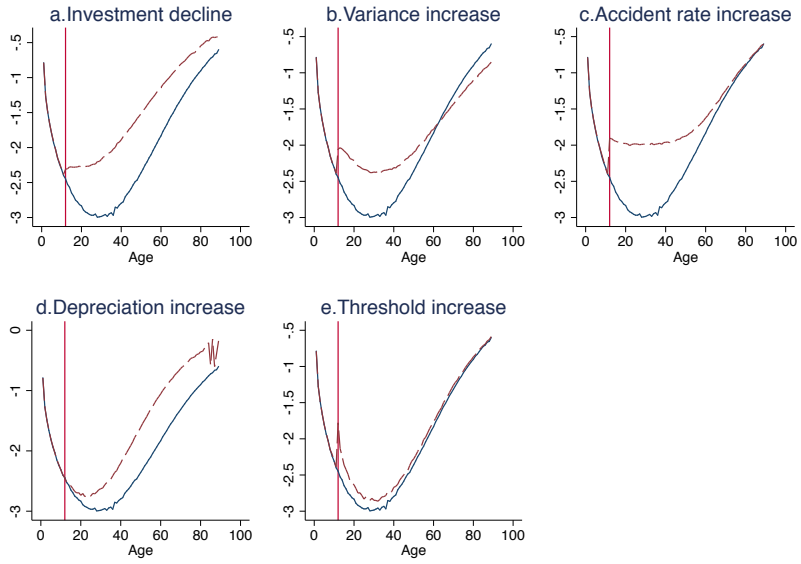
(b) Effects of decreasing annual health investments I throughout the lifetime by 50%



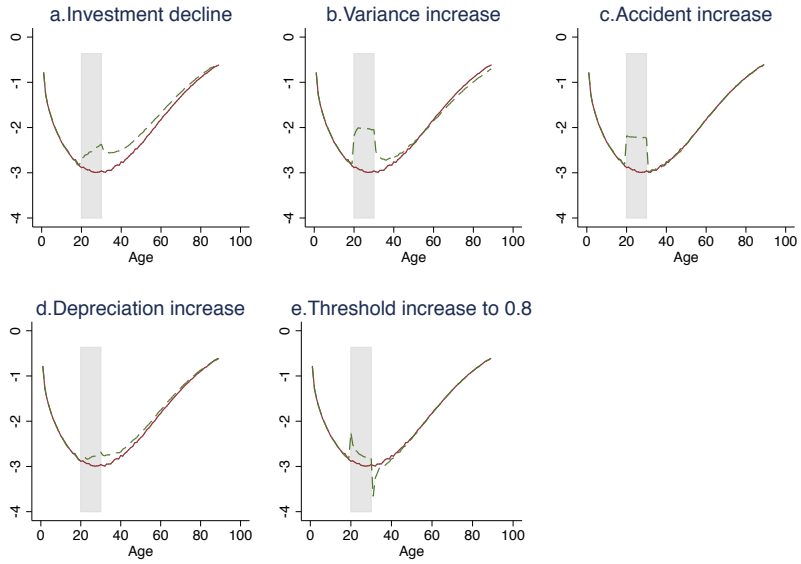
Note: Figure a and b show the gap in mortality and average health between two populations, that are identical except for one parameter. In The gap is computed as MR(I low)-MR(I high), or H(I low)-H(I high). The figures become very noisy after age 90 because there are almost no survivors, so we do not include these data points. Simulated data for two population of 500,000 individuals each. The baseline parameters are the same as in Figure 4.2b.

Figure 4.4: Effects of temporary and permanent shocks in adolescence

(a) Effect of permanent changes at age 12 on (the log of) mortality



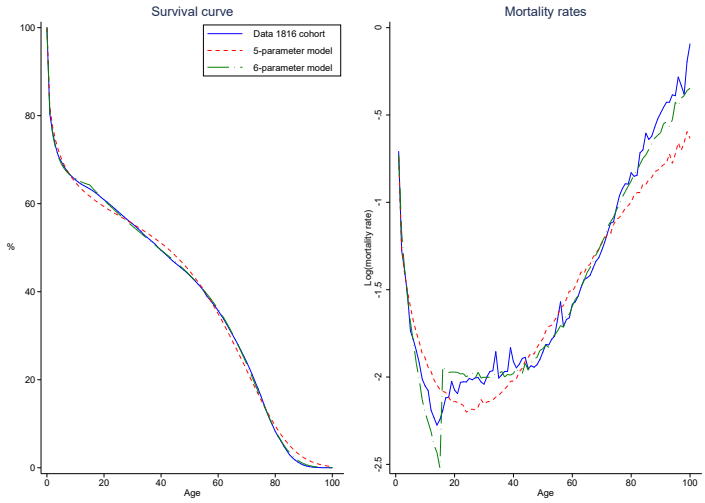
(b) Effect of exogenous temporary shocks at age 20 on log mortality



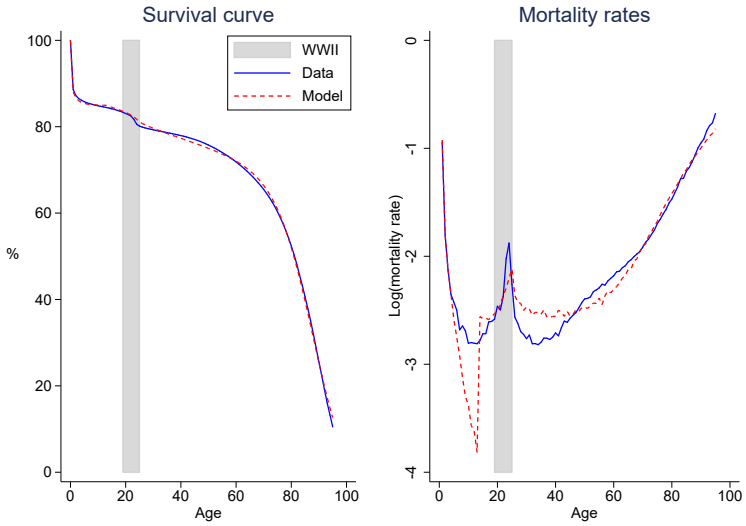
Note: the figures shows the effects of a permanent (Panel a) or temporary (panel b) change in a parameter occuring at age 12 (panel a) or at age 20 (Panel b) on the population's mortality. The solid line shows the profile of mortality for a baseline population and the dashed line shows the profile for the population who has a parameter change at age 12. Each populations has 500,000 individuals at birth. All increases (decreases) are 50% increases (decreases), except for the threshold which is increased to the level of μ. The shaded area in panel b corresponds to the years of the temporary shock. The baseline parameters are the same as in Figure 4.2b.

Figure 4.5: Model fit for two cohorts of French women

(a) Baseline estimates for the 1816 cohort, including adolescent hump



(b) WW2 as an investment shock. French Women born in 1921.



Note: In Figure a, the 5-parameter model assumes no accidents occur throughout the lifetime. The 6-parameter model allows for a positive accident rate to start in adolescence. See 4.1 for parameter values. The estimated parameters are shown in Appendix Table 4.3.

In Figure b, the model assumes a decrease in the investment level every year of the WWII. Mortality rates are shown in log(10) scale. The estimated parameters are shown in Appendix Table 4.4.

4.A Appendix A: Proofs omitted in the text

Before proving the propositions contained in the text we examine the identification of the model. The model is defined as follows:

$$\begin{cases} H_t = H_{t-1} - d(t) + I + \varepsilon_t & \text{if } D_{t-1} = 0 \\ D_t = \mathbb{I}(H_t \leq \underline{H}, D_{t-1} = 0), \\ D_0 = 0 \end{cases} \tag{4.1}$$

with $d(t) = \delta \cdot t^{\alpha}$ $\delta \in (0, \infty), \alpha \in (0, \infty),$ and $I \in \mathbb{R}$. $\underline{H}$ and σ_H^2 are normalized to be 0 and 1, respectively. Let $\hat{H}_t \equiv \mathbb{E}\left[H_t \mid H_t > 0\right]$ denote the average health in the living population with age t and $\sigma_{\hat{H}_t} \equiv Var\left[H_t \mid H_t > 0\right]$ the variance of health among the living.

Identification of the model Suppose we have two sets of parameters $\theta = (I, \delta, \sigma_\varepsilon, \alpha, \mu_H)$ and $\theta' = (I', \delta', \sigma'_\varepsilon, \alpha', \mu'_H)$. We say that θ and θ' are *observationally equivalent* (OE) if they imply the same mortality rates at each age, i.e. iff

$$MR_t(\theta) = MR_t(\theta'), \forall t \in \mathbb{N}$$

Equivalently, we could define observational equivalence in terms of survival rates $\{S_t(\theta)\}_{t>0}$ since each sequence can be uniquely recovered from the other one. Because we model does not have a closed-form solution for the mortality rates, proving the identification formally takes several steps. A key step is to use an "identification at infinity" argument to prove that the variances must be the same. Then identification of the remaining parameters follows.

We say that θ and θ' are *weakly observationally equivalent* (WOE) if and only if 1) θ and θ' are OE and 2) they do not generate the same sequences of health distributions, i.e.

$$\begin{cases} \exists t \in \mathbb{N}, \exists x \in \mathbb{R}^+ & F_{H_t}(x;\theta) \neq F_{H_t}(x;\theta') \\ \forall t \in \mathbb{N} & MR_t(\theta) = MR_t(\theta') \end{cases}$$

In constrast, we say that θ and θ' are *strongly observationally equivalent* (SOE) iff 1) θ and

θ' are OE and 2) they generate the same sequences of health distributions, i.e.

$$\begin{cases} \forall t \in \mathbb{N}, \quad F_{H_t}(\cdot\,;\theta) = F_{H_t}(\cdot\,;\theta') \\ \forall t \in \mathbb{N} \quad MR_t(\theta) = MR_t(\theta') \end{cases}$$

Although we cannot distinguish between OE and WOE bu observing mortality rate only, we could in principle observe some other features of the distributions of health at all ages that could break the identification.[39] Towards a contradiction, suppose that θ and θ' are OE, with $\theta' \neq \theta$. Then two cases must be considered, whether they are weakly or strongly observationally equivalent. We first show that strong observational equivalence is not possible.

Suppose that $(I, \delta, \alpha, \mu_H) \neq (I', \delta', \alpha', \mu'_H)$, then these two sets of parameters cannot generate the same first four modes of the health distribution and therefore are not observationally equivalent. The first four modes are

$$\begin{aligned} mode(1) &= \mu_H + I - \delta \\ mode(2) &= \mu_H + 2I - \delta(1 + 2^\alpha) \\ mode(3) &= \mu_H + 3I - \delta(1 + 2^\alpha + 3^\alpha) \\ mode(4) &= \mu_H + 4I - \delta(1 + 2^\alpha + 3^\alpha + 4^\alpha) \end{aligned} \tag{4.2}$$

These equations can be manipulated to obtain a triangular system: the difference between consecutive modes eliminate μ_H, $m_4 - m_3 = I - \delta 4^\alpha$, a double difference $(m_4 - m_3) - (m_3 - m_2)$ further eliminates I and finally dividing this double difference by another one, $m_3 - m_2 - m_2 - m_1$ eliminates δ. Now suppose that $(I, \delta, \alpha, \mu_H) = (I', \delta', \alpha', \mu'_H)$ and $\sigma_\varepsilon \neq \sigma'_\varepsilon$ then the first two mortality rates cannot be equal

$$m_1(\theta) = \Phi\left(\frac{-\mu_H - I + \delta}{\sqrt{1 + \sigma_\varepsilon^2}}\right) \neq \Phi\left(\frac{-\mu_H - I + \delta}{\sqrt{1 + (\sigma'_\varepsilon)^2}}\right) = m_1(\theta') \tag{4.3}$$

We now consider weak observational equivalence, i.e. the case where different sets of parameters produce different health distributions nonetheless delivering the same mortality rates. Let us define the first time at which the two health distribution differ $\tau \equiv$

[39] A step in that direction would be to observe a good proxy for health for a cohort.

$\min\{t \in \mathbb{N} \mid \exists x \in \mathbb{R}^+ \; F_{H_t}(x;\theta) \neq F_{H_t}(x;\theta')\}$. It is well defined by definition of weakly observational equivalence. Because F_{H_t} is continuous except at 0, we can assume without loss of generosity that the two cdf differ on some non trivial interval $(a,b) \supset \{x\}$. If $\tau > 1$ then because at the previous period the two distribution are the same, it must be the case that $\sigma_\varepsilon \neq \sigma'_\varepsilon$, otherwise the tails could not be similar. So it must be that $\tau = 1$ i.e. the distribution start differing at the first period.

Lemma 5. *The variance is separately identified.*

If the probability of surviving until age t , $\{S_t\}_{t\in[0,T]}$, is observed for an arbitrary large T. Then the variance is identified. Intuitively, the variance of the shock characterizes the thickness of the right-hand tail. If one population has a larger variance than the other one then the ratio of survivors grows arbitrarily large at old age. More formally, let $\theta = \{\sigma, \psi\}$ denote the set of parameters. For any ψ, ψ' such that

$$\sigma > \sigma' \Longrightarrow \lim_{t\to+\infty} \frac{S_t(\sigma,\psi)}{S_t(\sigma',\psi')} = +\infty \tag{4.4}$$

Let $\psi = \{\alpha, \delta, \mu_0, \kappa\}$ and $\sigma > \sigma'$. We have that for any t $\lim_{t\to+\infty} \frac{f_{H_t}(x;\sigma,\psi)}{f_{H_t}(x;\sigma',\psi')} = +\infty$. The only way to "compensate" for a small variance, which creates in old ages a right tail of very healthy people is to have a lower depreciation (δ and α). However because the tail decreases at exponential rate, we have

$$\lim_{t\to+\infty} \frac{f_{H_t}(x+z_t;\sigma,\psi)}{f_{H_t}(x;\sigma',\psi')} = +\infty \; \forall x > 0 \tag{4.5}$$

where $z_t = \sum_{s<t} \delta t^\alpha - \sum_{s<t} \delta' t^{\alpha'}$. This argument is essentially an "identification at infinity" (see Chamberlain 1986, Heckman 1990), a powerful tool to prove identification. For estimation purpose, however, this method is sensitive to the quality of the data sample. In our case, we verified by simulation that close values of σcan be difficult to distinguish as sampling error because larger for very old ages.

Lemma 6. *For any t, one of these cases occurs: either (1) f_{H_t} is hump-shaped (increasing then decreasing) or (2) f_{H_t} is strictly decreasing.*

The initial distribution and the shocks at every age are normal distributions, which are log-concave. The probability density function of a sum of two random variables is the convolution of their probability density functions. The space of log concave is closed under convolution (Ibragimov 1956) and Log-concavity implies single-peakedness. Log-concavity is not affected by the truncation either, until the point where the mode of the distribution at time t lies in the interval $(0, \delta(t+1)^{\alpha} - I)$. During the following period, the mode after convolution falls below 0 and the distribution becomes strictly decreasing henceforth.

Corollary 2. *There exists $t_{mode} > 0$ such that*

$$\begin{cases} mode\left(f_{H_t}\right) > 0 & t < t_{mode} \\ mode\left(f_{H_t}\right) = 0 & t \geq t_{mode} \end{cases}$$

where the mode of the distribution is

$$\max\left\{\mu_0 + I \cdot t - \delta \sum_{s=0}^{t} s^{\alpha}, 0^{+}\right\}$$

Proof of Proposition 24

1. Everyone dies eventually. The cumulative distribution function of our process can be bounded above by a process easier to study. Consider the process $\{H_t^*\}_{t=1}^{\infty}$, defined by $H_0^* = H_0 \sim \mathcal{N}(\mu_H, \sigma_H^2)$ and the recurrence relation:

$$H_t^* = H_{t-1}^* + I - \delta \cdot t^{\alpha} + \varepsilon_t \ , \ \varepsilon_t \sim \mathcal{N}\left(0, \sigma_{\varepsilon}^2\right) \tag{4.6}$$

The process is similar to the one in our model except that there is no truncation. It is easy to tell that $0 \leq P(H_t > z) \leq P(H_t^* > z)$ for any $z > 0$. Now for any $t \geq 0$, H_t^* is normally distributed with mean

$$\mu_{H_t^*} = \mu_H + I \cdot t - \delta \sum_{k=1}^{t} k^{\alpha} \tag{4.7}$$

and standard deviation

$$\sigma_{H_t^*} = \sqrt{\sigma_H^2 + t \cdot \sigma_{\varepsilon}^2} \tag{4.8}$$

Hence, $P\left(H_t^* > z\right) = 1 - \Phi(\frac{z-\mu_{H_t^*}}{\sigma_{H_t^*}})$, where Φ is the CDF of the standard normal distribution. As $t \to \infty$, we have $\mu_{H_t^*} \sim I \cdot t - \delta \cdot \frac{t^{\alpha+1}}{\alpha+1}$ and $\sigma_{H_t^*} \sim \sqrt{t} \cdot \sigma_\varepsilon$. Therefore if $\alpha > 0$, $\frac{\mu_{H_t^*}}{\sigma_{H_t^*}} \to -\infty$ as $t \to \infty$.

2. The mortality rates fall and then rise under suitable parametric restrictions.

The mortality rate at time t is defined as

$$MR_t = \frac{F_{H_t}(0) - F_{H_{t-1}}(0)}{1 - F_{H_{t-1}}(0)} \tag{4.9}$$

Consider the first period and suppose for a moment that $\sigma_\epsilon = 0$. For values of I, δ, α such that for instance $I - \delta(10)^\alpha > 0$, (i.e. the aging process does not turn negative until at least age 10), then there are no natural deaths until at least age 10 and $MR_2 \leq MR_1$. By continuity, as $\sigma_\varepsilon \to 0$ the number of accident deaths in the second period converges towards zero and therefore there is a neighborhood of 0 for which the number of deaths decline. Since the derivative of $x \mapsto \frac{1}{1-x}$ is equal to 1 at $x = 0$, the mortality rate declines if the numerator – the number of deaths – decreases. To show that the mortality rate rises, let's focus on the last periods when f_{H_t} is strictly decreasing, i.e. $t \geq t_1$ where $t_1, = \min\left\{t, \mu_0 + I \cdot t - \delta \sum_{s=0}^{t} s^\alpha < 0\right\}$. The accelerating aging pushes the distributions towards 0 at an increasing speed. At the same time because of the shock the right hand tail of F_{H_t} is fatter than the one of $F_{H_{t-1}}$ and therefore the successive distribution are not ordered by a stochastic dominance relation. To control the tail behavior, let $\eta > 0$ be arbitrary smal andl let $q$such that $F_{H_{t-1}}(q) = 1 - \eta$. The probability that those individuals on the right tail with $H_{t-1} > q$ died in the following period is bounded above by $\Phi\left(\frac{-q-I+\delta t^\alpha}{\sigma_\epsilon}\right)$.

3. The average health rises and then falls.

The average health of a cohort at age t is given by

$$\begin{aligned} \mathbb{E}\left[H_t \left| H_t > 0\right.\right] &= \mathbb{E}\left[H_{t-1} + I - \delta t^\alpha + \varepsilon_t \left| H_t > 0\right.\right] \\ &= I - \delta t^\alpha + \mathbb{E}\left[H_{t-1} \left| H_t > 0\right.\right] \\ &= I - \delta t^\alpha + \mathbb{E}\left[\mathbb{E}\left[H_{t-1} \left| H_{t-1} > 0\right.\right] \left| H_t > 0\right.\right] \end{aligned} \tag{4.10}$$

by the law of iterated expectations. Hence

$$\begin{aligned}\mathbb{E}\left[H_t \left| H_t > 0\right.\right] - \mathbb{E}\left[H_{t-1} \left| H_{t-1} > 0\right.\right] &= I - \delta t^{\alpha} \\ &+ \mathbb{E}\left[\mathbb{E}\left[H_{t-1} \left| H_{t-1} > 0\right.\right] \left| H_t > 0\right.\right] \\ &- \mathbb{E}\left[H_{t-1} \left| H_{t-1} > 0\right.\right]\end{aligned} \tag{4.11}$$

Notice that the second term is always positive because conditional on the individual surviving the next period, her health today must be above average as the shock are iid. As long as the second term can bounded as follows

$$0 \leq \mathbb{E}\left[\mathbb{E}\left[H_{t-1} \left| H_{t-1} > 0\right.\right] \left| H_t > 0\right.\right] - \mathbb{E}\left[H_{t-1} \left| H_{t-1} > 0\right.\right] \leq c + o(t^{\alpha}) \tag{4.12}$$

then we have the result (except in the middle region where $I - \delta t^{\alpha}$ is close to c) . To bound the second term, by definition, we have

$$\mathbb{E}\left[H_{t-1} \left| H_{t-1} > 0\right.\right] = \int_0^{\infty} x f_{H_{t-1}}(x)\, dx$$

Now,

$$\begin{aligned}P\left(H_{t-1} \leq x \mid H_t > 0\right) &= P\left(H_{t-1} \leq x \mid H_t > 0, H_{t-1} > 0\right) \\ &= \frac{P(H_{t-1} \leq x, H_t > 0 | H_{t-1} > 0)}{P(H_t > 0 | H_{t-1} > 0)} \\ &= \frac{1 - F_{H_{t-1}}(0)}{1 - F_{H_t}(0)} \int_0^x f_{H_{t-1}}(u) \left[1 - \Phi\left(\delta t^{\alpha} - I - u\right)\right] du\end{aligned} \tag{4.13}$$

hence

$$\mathbb{E}\left[H_{t-1} \left| H_t > 0\right.\right] = \int_0^{\infty} x f_{H_{t-1}}(x) \left\{ \frac{1 - F_{H_{t-1}}(0)}{1 - F_{H_t}(0)} \left[1 - \Phi\left(\delta t^{\alpha} - I - x\right)\right] \right\} dx \tag{4.14}$$

and finally

$$\begin{aligned}&\mathbb{E}\left[\mathbb{E}\left[H_{t-1} \left| H_{t-1} > 0\right.\right] \left| H_t > 0\right.\right] - \mathbb{E}\left[H_{t-1} \left| H_{t-1} > 0\right.\right] \\ &= \int_0^{\infty} x f_{H_{t-1}}(x) \left\{ \frac{1 - F_{H_{t-1}}(0)}{1 - F_{H_t}(0)} \left[1 - \Phi\left(\delta t^{\alpha} - I - x\right)\right] - 1 \right\} dx\end{aligned} \tag{4.15}$$

Fix $\epsilon > 0$, for any t there is an $\underline{x}$ such that $x > \underline{x} \Rightarrow \Phi\left(\delta t^{\alpha} - I - x\right) < \epsilon$. For large values of x the term in square bracket goes to 0 very fast. In addition because $f_{H_{t-1}}(.)$ decreases at an

exponential rate, we can bound the right tail by ϵ of the integrand and focus on integrating from 0 to $\underline{x}$. As $t \to \infty$ this term goes to 0. Therefore ultimately the average health decreases.

4. The Variance rises and then falls to 0

There are two forces i) adding the shock always spreads out the distribution and increases the variance by an amount σ_ϵ, ii) the truncation always reduces the variance. Because the force of aging initially pushes the distribution away from the threshold, it reduces the impact of the truncation. Once the sign of $I-\delta t^\alpha$ turns negative the compression towards the threshold becomes dominant. initially as we add to the health stock an iid noise. We have $Var\left(H_t\right) = Var\left(H_{t-1}\left|D_{t-1}=0\right.\right)+\sigma_\epsilon$. Because of the truncation we have $Var\left(H_t\left|D_t=0\right.\right) \leq Var\left(H_t\right)$. To bound the gap between the two quantities, and show that is small compared to σ_ϵ initially, we need to bound the contribution of the left tail to the variance. To bound the gap, notice that $\mathbb{E}\left[\left(\mathbb{E}\left[H_t\left|H_t<0\right.\right]-\mathbb{E}\left[H_t\left|H_t>0\right.\right]\right)^2\right] \leq \int_{F^{-1}(MR_t)}^{+\infty}\left(y+\mathbb{E}\left[H_t\left|H_t>0\right.\right]\right)^2\phi\left(\frac{y}{\sigma_\epsilon}\right)dy$, where the right hand side corresponds to mean square distance with the mean of health at time t of a mass of individuals starting at $H_{t-1}=0$ and hit by a negative shock. The right hand side can be computed since it corresponds upt to a change of variable to the second moment of a truncated normal. Ultimately, it is clear that the variance converges to 0, since for all $x>0$, $\lim_{t\to\infty} F_{H_t}\left(x\right)=0$.

Proof of Proposition 25

1. Increasing the investment I: Let $a_t = I-\delta t^\alpha$. The random variable H_t has a mass point at $z=0$ but is continuous on $(0,+\infty)$. $F_{H_t}\left(0\right)$ is the probability of not surviving until age t while for any $z>0$, the cdf can be expressed

$$F_{H_t}\left(z\right)=\int_{x=0}^{\infty}\Phi\left(\frac{z-a_t-x}{\sigma_\varepsilon}\right)f_{H_{t-1}}\left(x\right)dx+F_{H_{t-1}}\left(0\right) \tag{4.16}$$

Equivalently, after integration by parts, one obtains:

$$F_{H_t}\left(z\right)=-\frac{1}{\sigma_\varepsilon}\int_{x=0}^{\infty}\phi\left(\frac{z-a_t-x}{\sigma_\varepsilon}\right)F_{H_{t-1}}\left(x\right)dx+F_{H_{t-1}}\left(0\right) \tag{4.17}$$

Hence the mortality rate at age t, which is the probability of dying at age t conditional on surviving until age t, can be written:

$$MR_t = \frac{F_{H_t}(0) - F_{H_{t-1}}(0)}{1 - F_{H_{t-1}}(0)}$$

Suppose that for every t we increase the constant investment level I to some level $I' > I$. Following the expression above, the impact can be be decomposed in two: first, a direct effect on the probability of dying at age t (the numerator) and, second, a compounded effect carried through the distribution of health for those attaining age t.We show that, for any t, both effects go in the same direction: an increase in I simultaneously increases the probability of surviving untill age t (hence increases the denominator) and reduces the probability of dying at age t (the numerator goes down). We prove the following lemma.

Lemma 7. *For all t, we have:*

1. $\forall z \geqslant 0$, $\frac{\partial F_{H_t}(z;I)}{\partial I} \leq 0$
2. $\frac{\partial MR_t}{\partial I} \leq 0$

Proof: We prove these inequalities jointly and by induction. Notice that $\frac{\partial F_{H_t}(\cdot;I)}{\partial I} \leq 0$ signifies that the cdf's are ranked by first order stochastic dominance. The higher the I, the further the distribution is pushed to the right, which decreases the value of the cdf at any point x as I increases. Because all the individuals would then be in better health, ceteris paribus, fewer of them will die each period. Combined with a higher denominator, this delivers a lower mortality rate at each point.

At $t = 0$: $F_{H_1}(z;\, I) = \Phi\left(\frac{z-\mu_0}{\sigma_0}\right)$ hence $\frac{\partial F_{H_1}(z;I)}{\partial I} = 0$. $MR_t = F_{H_t}(0) = \Phi\left(\frac{z-\mu_0}{\sigma_0}\right)$ which, again, is non-increasing with I. In period 1, we have $\forall z \geqslant 0$,

$$\begin{aligned} F_{H_1}(z) &= Pr(H_0 + I - \delta + \varepsilon_1 \leqslant z) \\ &= Pr(H_0 + \varepsilon_1 \leqslant z + \delta - I) \\ &= \Phi\left(\frac{z + \delta - I - \mu_H}{\sqrt{1 + \sigma_\varepsilon^2}}\right) \end{aligned} \tag{4.18}$$

therefore in terms of mortality rates we have

$$\begin{aligned} MR_1 &= Pr(H_1 \leqslant 0) \\ &= Pr(H_0 + I - \delta + \varepsilon_1 \leqslant 0) \\ &= \Phi\left(\frac{\delta - I - \mu_H}{\sqrt{1+\sigma_\varepsilon^2}}\right) \end{aligned} \tag{4.19}$$

It is easy to see that as I or μ_H increases, both $F_{H_1}(z)$ and MR_1 decreases. We have, $\forall t \geqslant 2$, $\forall z \geqslant 0$,

$$\begin{aligned} F_{H_t}(z) &= Pr(H_t \leqslant z) \\ &= Pr(H_t \leqslant z, H_{t-1} > 0) + Pr(H_t \leqslant z, H_{t-1} \leqslant 0) \\ &= Pr(H_{t-1} + I - \delta t^\alpha + \varepsilon_t \leqslant z, H_{t-1} > 0) + Pr(H_{t-1} \leqslant 0) \\ &= Pr(0 < H_{t-1} \leqslant z + \delta t^\alpha - I - \varepsilon_t) + Pr(H_{t-1} \leqslant 0) \\ &= \frac{1}{\sigma_\varepsilon}\int_{x=0}^{\infty} \phi\left(\frac{x - z - \delta t^\alpha + I}{\sigma_\varepsilon}\right)\left(F_{H_{t-1}}(x) - F_{H_{t-1}}(0)\right)dx + F_{H_{t-1}}(0) \end{aligned} \tag{4.20}$$

$$MR_t = \frac{F_{H_t}(0) - F_{H_{t-1}}(0)}{1 - F_{H_{t-1}}(0)} \tag{4.21}$$

Then

$$\begin{aligned} &\frac{\partial F_{H_t}(z;I)}{\partial I} \\ =&\frac{\partial}{\partial I}\left(\frac{1}{\sigma_\varepsilon}\int_{x=0}^{\infty} \phi\left(\frac{x - z - \delta t^\alpha + I}{\sigma_\varepsilon}\right)\left(F_{H_{t-1}}(x) - F_{H_{t-1}}(0)\right)dx\right) + \frac{\partial F_{H_{t-1}}(0;I)}{\partial I} \\ =&\frac{1}{\sigma_\varepsilon^2}\int_{x=0}^{\infty} \phi'\left(\frac{x - z - \delta t^\alpha + I}{\sigma_\varepsilon}\right)\left(F_{H_{t-1}}(x) - F_{H_{t-1}}(0)\right)dx \\ &+ \frac{1}{\sigma_\varepsilon}\int_{x=0}^{\infty} \phi\left(\frac{x - z - \delta t^\alpha + I}{\sigma_\varepsilon}\right)\left(\frac{\partial F_{H_{t-1}}(x;I)}{\partial I} - \frac{\partial F_{H_{t-1}}(0;I)}{\partial I}\right)dx \\ &+ \frac{\partial F_{H_{t-1}}(0;I)}{\partial I} \\ =&\frac{1}{\sigma_\varepsilon^2}\int_{x=0}^{\infty} \phi'\left(\frac{x - z - \delta t^\alpha + I}{\sigma_\varepsilon}\right)\left(F_{H_{t-1}}(x) - F_{H_{t-1}}(0)\right)dx \\ &+ \frac{1}{\sigma_\varepsilon}\int_{x=0}^{\infty} \phi\left(\frac{x - z - \delta t^\alpha + I}{\sigma_\varepsilon}\right)\frac{\partial F_{H_{t-1}}(x;I)}{\partial I}dx \\ &+ \Phi\left(-z - \delta t^\alpha + I\right)\frac{\partial F_{H_{t-1}}(0;I)}{\partial I} \end{aligned} \tag{4.22}$$

All three items are negative, so we have

$$\frac{\partial F_{H_t}(z;I)}{\partial I} \leqslant 0 \tag{4.23}$$

Differentiating the mortality rates with respect to investments, we have

$$\begin{aligned}
&\frac{\partial MR_t}{\partial I} \\
=&\frac{1}{(1-F_{t-1}(0))^2}\left(\frac{\partial F_{H_t}(0)}{\partial I}(1-F_{H_{t-1}}(0)) - \frac{\partial F_{H_{t-1}}(0)}{\partial I}(1-F_{H_t}(0))\right) \\
\leq&0
\end{aligned} \tag{4.24}$$

2. Increasing any of the aging parameters, δ or α

The exact same proof applies for δ and α as their impact on F_{H_t} through the aging function a_t , is similar to the effect of I.

3. An increase in σ_H^2 is ambiguous. However, it can be seen right away that this proof will not work for σ_ε nor σ_0. Increasing any of these variances, – a mean-preserving spread – will *not* give rise to the first order stochastic ranking of the cdf's that we have used. Increasing MR on impact. The numerator of the MR_t is given by

$$F_{H_t}(0;\, I) - F_{H_{t-1}}(0;\, I) = \int_{x=0}^{\infty} \Phi\left(\frac{\delta t^\alpha - I - x}{\sigma_\varepsilon}\right) f_{H_{t-1}}(x)\, dx$$

Since Φ is nondecreasing, if one decreases σ_ε at time t, and at this period only, then this expression is necessarily decreasing in σ_ε if $\delta t^\alpha \leq I$. This follows from the fact that a higher $\sigma_{\varepsilon,t}$ will generate a fatter right-hand tail. For instance, $\lim_{x\to+\infty} \frac{f_{H_t}(x;\sigma_\varepsilon)}{f_{H_t}(x;\sigma'_\varepsilon)} = 0$. Now if $\sigma_{\varepsilon,t}$ is changed only at period . From then on, the distributions are modified through a similar process. It can be shown that the fatter right-hand tail property will be preserved. In the very old age, only the popupulation in the right-tail have survived, hence the result. That the fatter right-hand tail property is preserved, is proved similarly by inference.

Remark 1: Lemma 1 is actually a subcase of the following result, which is slightly stronger: Suppose that the level of investment is allowed to change at every period, and denote $\mathcal{I} =$

$\{I_1, I_2, ..., \}$ and $\mathcal{I}' = \{I'_1, I'_2, ..., \}$ two investment sequences. The following holds:

$$\forall s \geq 1,\ I'_s \geq I_s \Longrightarrow \forall t \geq 1, \forall z > 0, F_{H_t}(z; \mathcal{I}) \leq F_{H_t}(z; \mathcal{I}') \text{ and } MR_t(\mathcal{I}) \leq MR_t(\mathcal{I}')$$

The mechanics of the proof is almost exactly similar. Increasing investment at any period generates a persistent relation of first-order stochastic dominance in the CDF of health.

4. Investment and health at birth are complements The proof is similar to the proof of 1. in Proposition 1

$$\begin{array}{rcl} \frac{\partial^2 F_{H_{t_2}}(z;\mathcal{I})}{\partial I_{t_1} \partial I_{t_2}} & = & \frac{\partial}{\partial I_{t_1}} \frac{\partial}{\partial I_{t_2}} \left[-\frac{1}{\sigma_\varepsilon} \int_{x=0}^{\infty} \phi\left(\frac{z-x-I_{t_2}+\delta(t_2)^\alpha}{\sigma_\varepsilon} \right) \left(F_{H_{t_2-1}}\left(x, \mathcal{I} - F_{H_{t_2-1}}(0)\right)\right) dx \right] \\ & + & \frac{\partial}{\partial I_{t_1}} \frac{\partial F_{H_{t_2-1}}(0;\mathcal{I})}{\partial I_2} \\ & = & \frac{\partial}{\partial I_{t_1}} \left[\frac{1}{\sigma_\varepsilon^2} \int_{x=0}^{\infty} \phi'\left(\frac{z-x-a_t}{\sigma_\varepsilon} \right) \left(F_{H_{t_2-1}}(x, \mathcal{I}) - F_{H_{t_2-1}}(0)\right) dx \right] + 0 \\ & = & \frac{1}{\sigma_\varepsilon^2} \int_{x=0}^{\infty} \phi'\left(\frac{z-x-a_t}{\sigma_\varepsilon} \right) \left(\frac{\partial}{\partial I_{t_1}} F_{H_{t_2-1}}(x, \mathcal{I}) - \frac{\partial}{\partial I_{t_1}} F_{H_{t_2-1}}(0)\right) dx \\ & \leq & 0 \end{array} \tag{4.25}$$

because $\frac{\partial}{\partial I_{t_1}} F_{H_{t_2-1}}(x, \mathcal{I}) \leq 0$ (increasing investment at time 1 creates a FOSD distribution). And as a consequence the denominator $1 - F_{H_{t_2-1}}(0)$ goes up as well.

Remark: Extended model with Accident shocks Proposition 1 and Proposition 2 hold for the extended model with accident shocks drawn indepently from the health status. This is because of independence, the accident shock leaves the health distributions unchanged and the proofs are unaffected.

4.B Supplementary Tables and Figures

Table 4.1: Modeling prime-age mortality. French Women born in 1816

		(0)	(1)	(2)	(3)	(4)
Model for hump: change in...		Baseline	I	$\underline{H}$	σ_e	κ_a
Initial mean health	μ_H	0.9115	0.9115	0.8151	0.7723	0.8634
Investment	I	0.1336	0.1336	0.1315	0.2159	0.4075
Standard Deviation of Shock	σ_e	0.5556	0.5556	0.4830	0.6300	1.0241
Depreciation	δ	0.0010	0.0010	0.0008	0.0009	0.0006
Aging	α	1.4350	1.4350	1.4462	1.5605	1.7849
Adolescent Hump*			0.1336	0.5586	0.9024	0.0086
Fit (survival curve)^		155.06	155.06	123.03	97.04	12.36
Fit (log of q_x)		3.01	3.01	2.87	2.23	0.74
Fit (death distribution)**		6.21	6.21	4.95	2.95	3.35
Actual Life Expectancy				38.25		
Predicted Life Expectancy		38.43	38.43	38.38	38.45	38.28
Counterfactual Life expectancy^^			38.43	40.92	40.38	45.86

*The estimate in this row corresponds to the value of the parameters after the onset of adolescence. Adolescence starts at age = (- 0.0175 x calendar year) + 47.4 for all women, based on the estimates provided in La Rochebrochard 2000.

**To make the fit of the age distribution comparable across columns we use the (normalized) number of deaths as weights.

^Our main fit criteria is the sum of squared errors of the survival rate at each age We also report the fit as the sum of squared errors of the log of q_x (the probability of dying between ages x and $x+1$) and the distribution of deaths. We don't target these moments directly–we target the survival curve.

^^Counterfactual Life Expectancy is computed by holding all estimated parameters fixed and setting the adolescent hump to 0.

Table 4.2: Estimated parameters for female chimpanzees living in the wild

Gender		Basic model	κ_a
Initial mean health	μ_H	0.9783	1.0043
Investment (annual)	I	0.3295	0.3390
Standard Deviation of Shock	σ_e	1.0871	1.1304
Depreciation	δ	0.0560	0.0553
Aging	α	0.7677	0.7820
Adolescent Hump*	κ_a		0.00001
# of individuals at birth		80	80
# of moments reported		55	55
Fit (survival curve)^b		112.50	111.29
Fit (log of q_x)		2.11	2.11
Actual Life Expectancy		15.38(13.4)^a	
Predicted Life Expectancy		15.35	15.35

Data sources: Life tables for primates in the wild come from Bronikowski et al. 2011. In the wild population data come from Brazil, Costa Rica, Kenya, Tanzania, Madagascar and Rwanda.

a. Life expectancy in parenthesis corresponds to the one reported in Bronikowski et al. 2011.

b. We target the survival curve and compute the sum of squared errors – the data provided are in the form of survival rates.

*Adolescence starts at age 8.

Table 4.3: Modeling prime-age mortality French Women born in 1860

		(0)	(1)	(2)	(3)	(4)
Model for hump: change in...		Baseline	I	$\underline{H}$	σ_e	κ_a
Initial mean health	μ_H	1.0981	1.0740	1.0748	0.9589	0.9323
Investment	I	0.1501	0.1563	0.1649	0.4879	0.3318
Standard Deviation of Shock	σ_e	0.5916	0.5873	0.5907	1.0471	0.7932
Depreciation	δ	0.0006	0.0005	0.0006	0.0001	0.0004
Aging	α	1.5742	1.5774	1.5889	2.2001	1.7780
Adolescent Hump*	κ_a		0.1380	0.4902	2.1746	0.0071
Fit (survival curve)^		205.54	197.74	177.41	46.34	11.93
Fit (log of q_x)		2.59	2.49	2.57	1.62	0.70
Fit (death distribution)**		7.49	19.78	23.43	9.07	16.48
Actual Life Expectancy				43.80		
Predicted Life Expectancy		43.95	44.04	43.97	43.88	43.85
Counterfactual Life expectancy^^			46.30	45.91	48.96	51.65

*The estimate in this row corresponds to the value of the parameter after the onset of adolescence. Adolescence starts at age = (- 0.0175 x calendar year) + 47.4 for all women except for the 5th column where the timing of adolescence is estimated as following a normal distribution with mean value (- 0.0175 x calendar year) + 47.4, and standard deviation 1.3285 (calculated from the table of 1975 girls) based on the estimates provided in La Rochebrochard 2000.

**To make the fit of the age distribution comparable across columns we use the (normalized) number of deaths as weights.

^Our main fit criteria is the sum of squared errors of the survival rate at each age We also report the fit as the sum of squared errors of the log of q_x (the probability of dying between ages x and $x+1$) and the distribution of deaths. We don't target these moments directly–we target the survival curve.

^^Counterfactual Life Expectancy is computed by holding all estimated parameters fixed and setting the adolescent hump to 0.

Table 4.4: Estimated parameters for WWII for French Women born in 1921

		(1)	(2)	(3)	(4)	(5)
Model for WWII: change in...		I	κ_a	σ_e	$\underline{H}$	δ
Initial condition	μ_H	0.9790	1.0837	1.0638	1.0522	1.0875
Investment	I	0.2985	0.2739	0.2650	0.2385	0.2597
Standard Deviation of Shock	σ_e	0.4255	0.4561	0.4358	0.3891	0.4412
Depreciation	δ	0.0007	0.0008	0.0009	0.0008	0.0008
Aging	α	1.5358	1.5272	1.4961	1.4785	1.5121
Adolescence Hump*	κ_a	0.0026	0.0030	0.0032	0.0031	0.0032
WWII Shock**		-0.1173	0.0036	0.3495	0.8919	0.0000
Fit (survival curve)^		40.62	37.02	38.49	38.53	38.88
Fit (log of q_x)		4.87	2.69	3.11	3.29	2.87
Fit during WWII (log of q_x)		0.21	0.53	0.68	0.73	0.68
% Difference in # deaths during WWII~~		-0.21	-0.36	-0.45	-0.32	-0.45
Actual Life Expectancy				66.00		
Predicted Life Expectancy		66.03	66.03	66.02	66.03	66.02
Counterfactual Life expectancy^^		70.90	66.24	65.94	66.23	65.05
Actual Life expectancy in 1946				55.93		
Life expectancy in 1946		55.27	55.25	55.12	55.28	55.16
Counterfactual LE in 1946		60.36	55.24	55.05	55.17	54.00

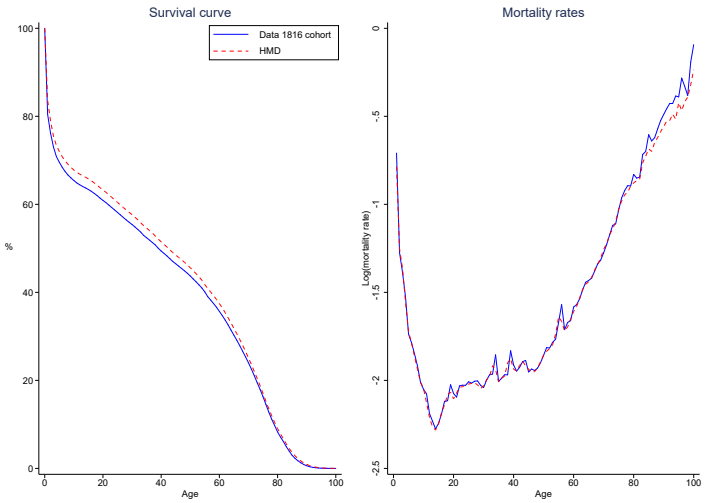
*Hump is modeled as a accident rate that starts in adolescence, set to happen at (- 0.0175 * calendar year) + 47.4 based on the estimates provided in La Rochebrochard 2000.

**The estimates in this row corresponds to the value of the parameter during the war. For example the first column shows that I was about 0.299 throughout life but decreased to -0.117 during the war. The same applies to columns (3) and (4), the standard deviation decreases from 0.436 to 0.350 and the threshold moves from 0 to 0.892. In column 2, we estimate the value of an additional random shock during the war, an approximate 41% decrease relative to the adolescent hump (but since the shock is independent this is only approximate).

^Our main fit criteria is the sum of squared errors of the survival rate at each age We also report the fit as the sum of squared errors of the log of q_x (the probability of dying between ages x and $x+1$). We don't target these moments directly–we target the survival curve.

^^Counterfactual Life Expectancy is computed by holding all estimated parameters fixed and setting the war parameters to 0

Figure 4.6: Comparison of q-rate in the paper and HMD (1816)



Life expectancy: 38.25 (with the q we use) and 39.86 (with the q in HMD). The life expectancy in HMD is 39.83.

Figure 4.7: Age profile of mortality of women born in France between 1860 and 1940, by decade

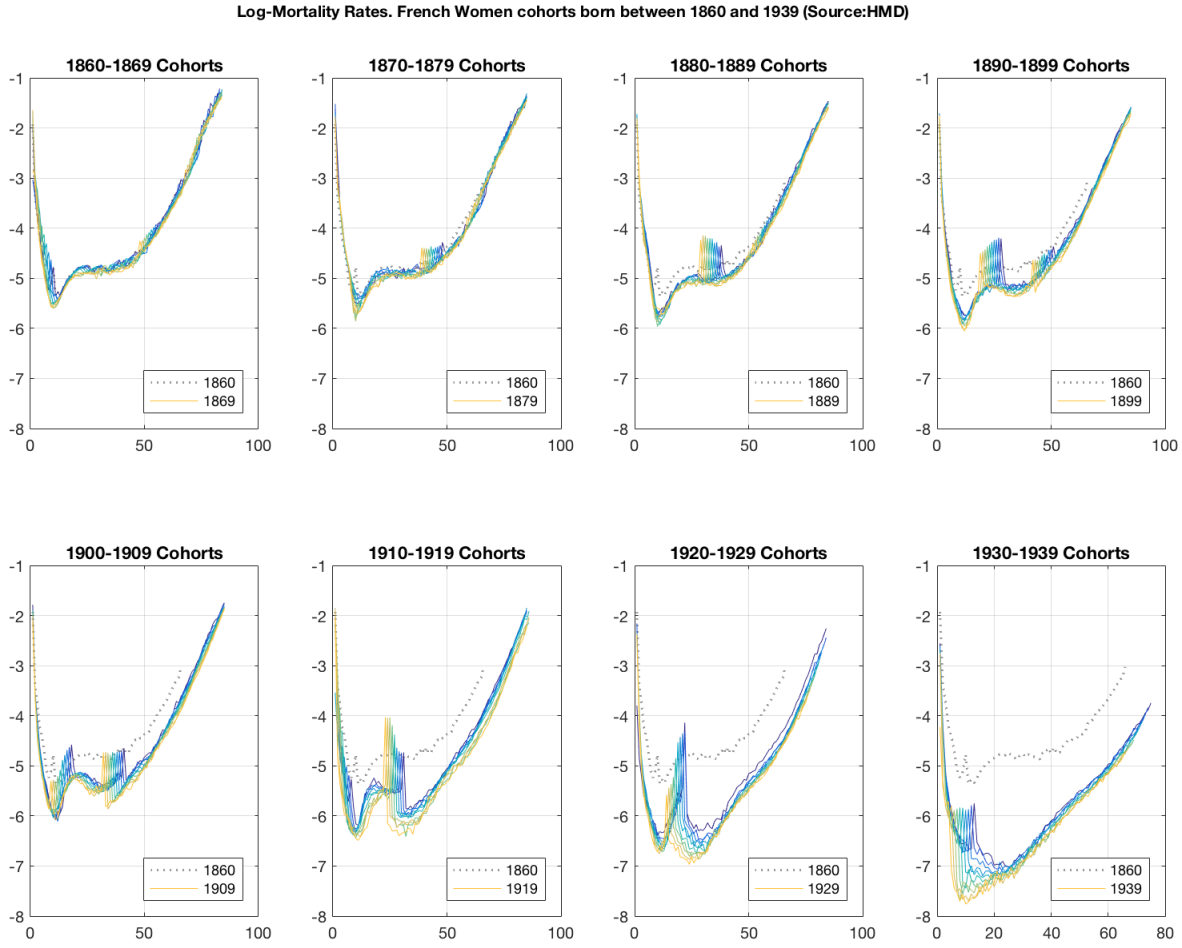
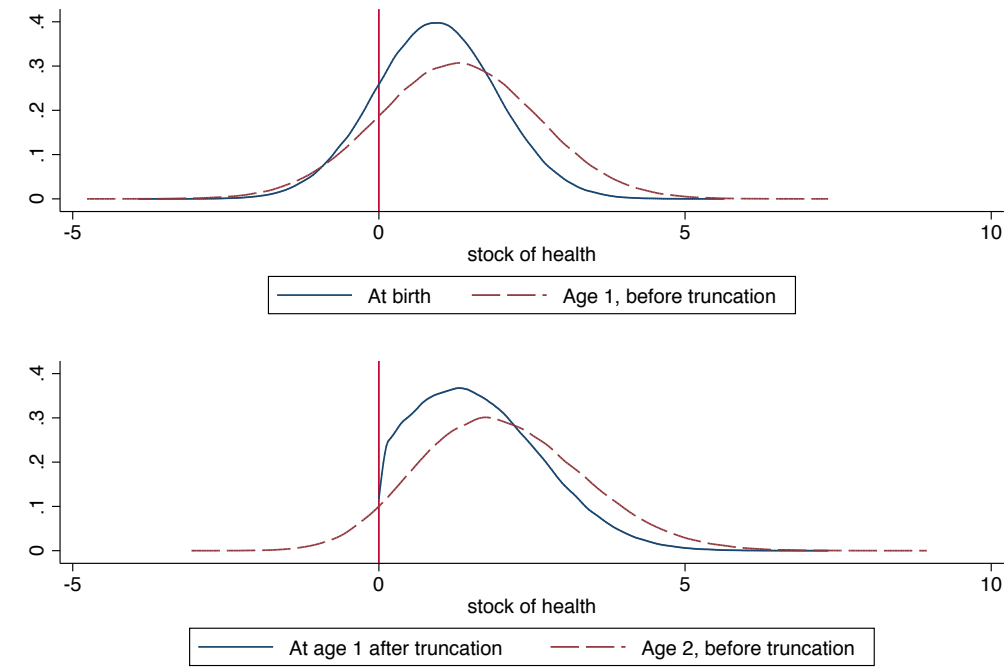


Figure 4.8: Health and mortality in the first two years of life



Data from simulations

In the first period the (infant) mortality rate MR_1 is given by

$$MR_1 = P(H_1 \leq \underline{H}) = P\left(H_0 + I - \delta + \varepsilon_1 \leq \underline{H}\right)$$

$$= P(\varepsilon_1 \leq \varphi_1) = F(\varphi_1)$$

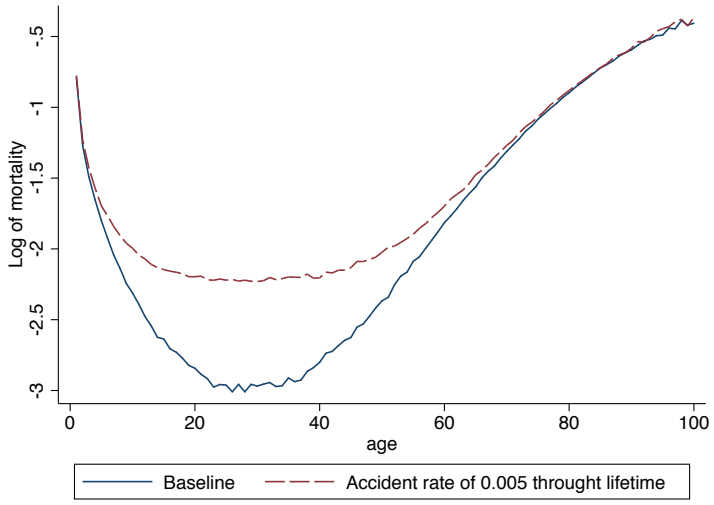
where $\varphi_1 = \underline{H} - I + \delta - H_0$ captures the threshold for dying in period 1 in terms of the random shock. Investments lower this threshold (lower mortality) and depreciation increases it (increases mortality).

Consider now the probability of dying at age $t = 2$. This is given by the probability that the stock falls below $\underline{H}$ at age 2, conditional on having survived to age 2, which can be expressed as:

$$\begin{aligned} MR_2 &= E(D_2 = 1 | D_1 = 0) = P(H_2 < \underline{H} | H_1 > \underline{H}) \\ &= \frac{P(H_2 < \underline{H}, H_1 > \underline{H})}{P(H_1 > \underline{H} | g_1, g_2)} = \frac{P(\varepsilon_2 < \varphi_2 - \varepsilon_1, \varepsilon_1 > \varphi_1)}{1 - F(\varphi_1)} \\ &= \frac{K(\varphi_2, \varphi_1)}{1 - F(\varphi_1)} \end{aligned} \tag{4.26}$$

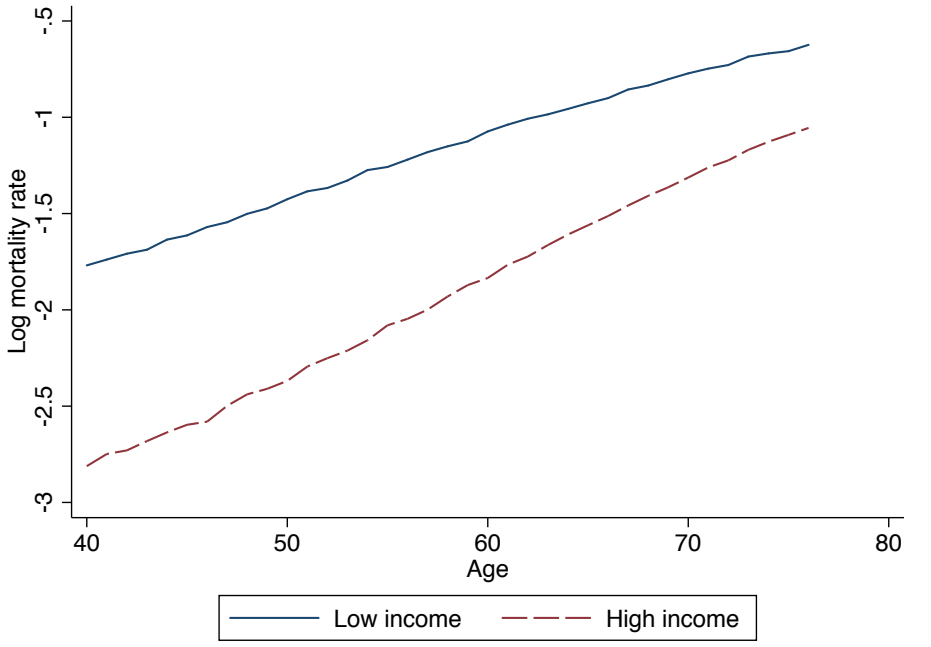
where $\varphi_2 = \underline{H} - I + \delta 2^{\alpha} - H_0$ captures the threshold for dying in period 2, and $K(\varphi_2, \varphi_1) = \int_{\varepsilon_1=\varphi_1}^{\infty} \int_{\varepsilon_2=-\infty}^{\varphi_2-\varepsilon_1} f(\varepsilon_1) f(\varepsilon_2) d\varepsilon_1 d\varepsilon_2$ is the density right above the old threshold and below the new threshold, that is the fraction of survivors who dies as a result of a new shock. The denominator is the fraction of survivors.

Figure 4.9: Adding accidents to the baseline model



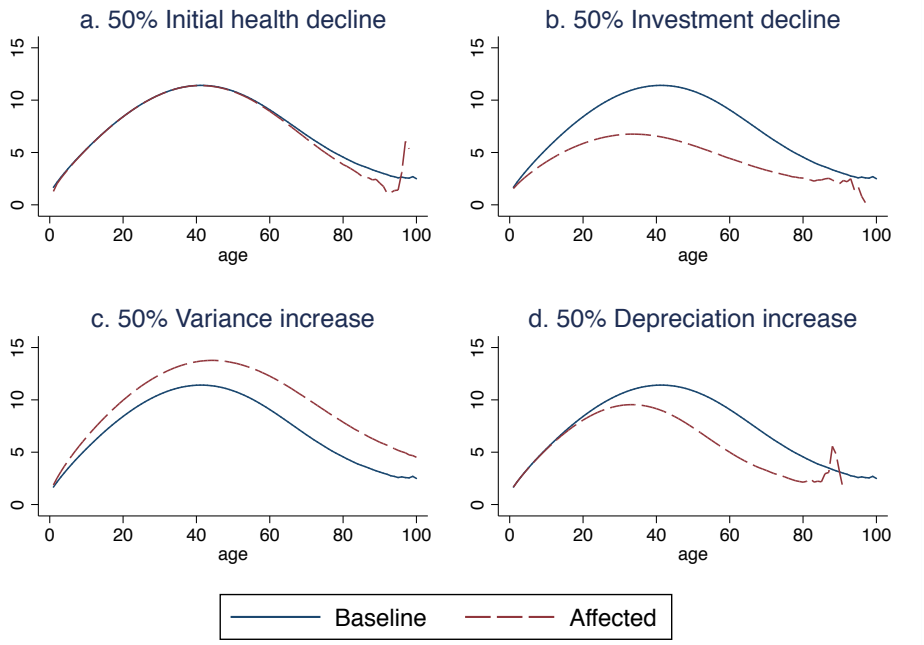
Note: The baseline parameters are the same as in Figure 4.2b

Figure 4.10: Comparative statics for log mortality



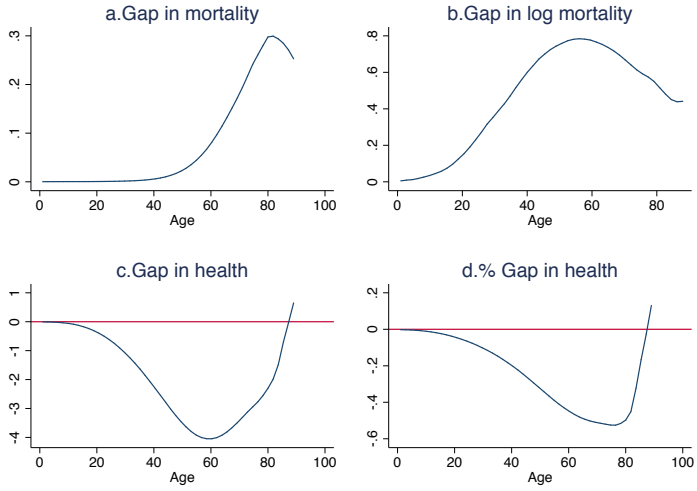
Note: Simulated data for two population of 500,000 individuals each. The baseline parameters are the same as in Figure 4.2b

Figure 4.11: Comparative statics for health



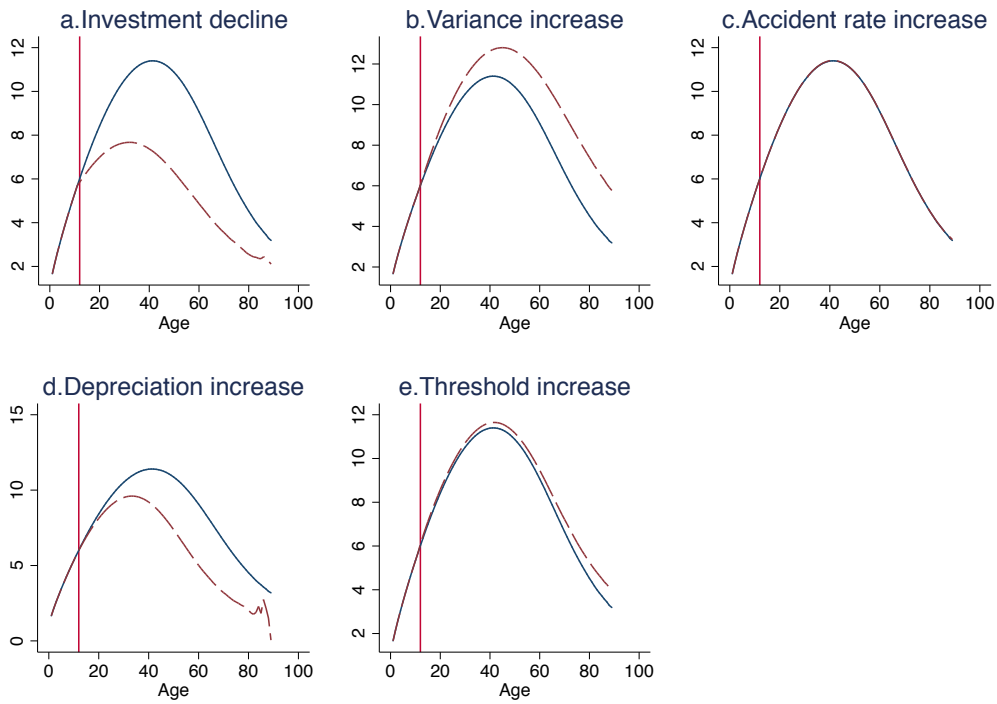
Note: Simulated data for a population of 500,000 individuals. The figures show the effect of changes relative to the baseline model, which is simulated using the same parameters we used for Figure 2.

Figure 4.12: Increasing the lifetime depreciation rate by 50% by age



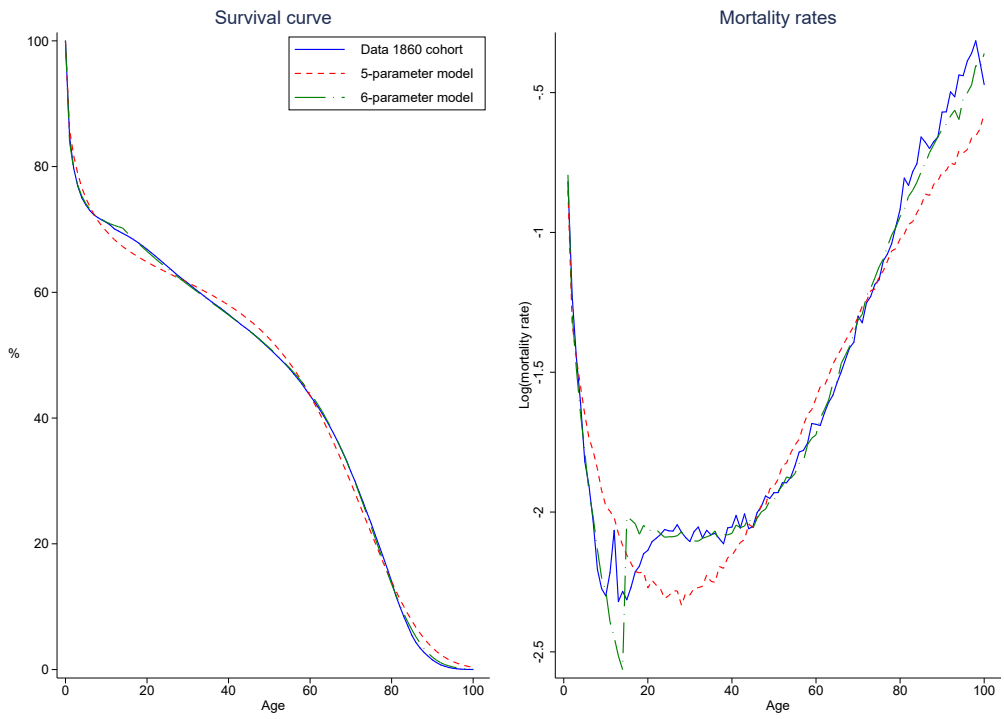
Note: The Figure shows the gap in mortality or health between a basline population and a population with a 50% higher depreciation rate. Gap is computed as MR(low)-MR(high), or H(low)-H(high). The figures become very noisy after age 90 because there are almost no survivors, so we do not include these data points. Simulated data for two population of 500,000 individuals each. The baseline parameters are the same as in Figure 4.2b

Figure 4.13: Health Effects of permanent shocks at age 12



Note: the figure shows the effects of a permanent change in a parameter occuring at age 12 on the average population health. The solid line shows the profile of average health for a baseline population and the dashed line shows the profile for the population who has a parameter change.

Figure 4.14: Baseline model with adolescent hump, 1860 cohort



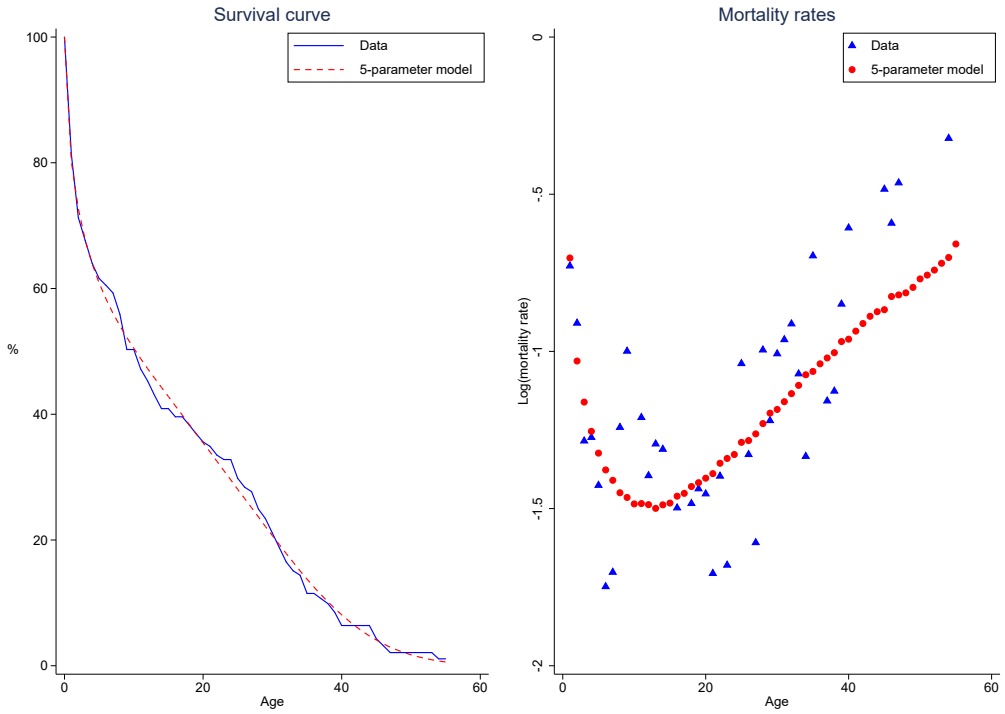
Note: Mortality rates are shown in log(10) scale. The estimated parameters are in Appendix Table 4.14.

4.C Notes on the empirical method

1. Data

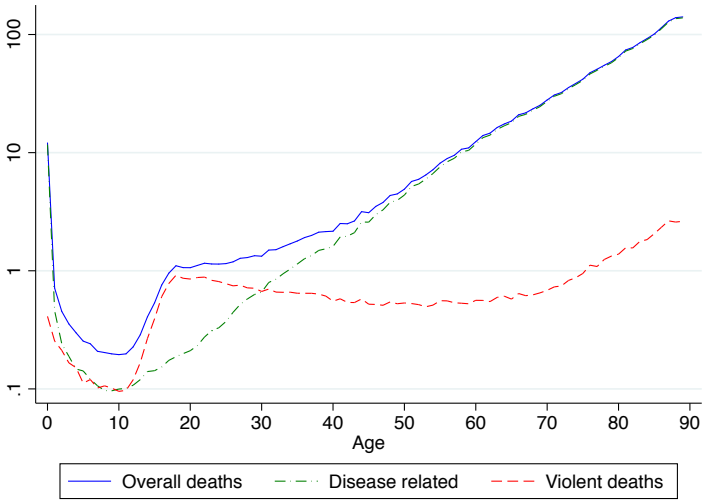
Territory changes. The table below describes the details of the changes in territory that took place in France since 1816.

Figure 4.15: Survival curve for apes



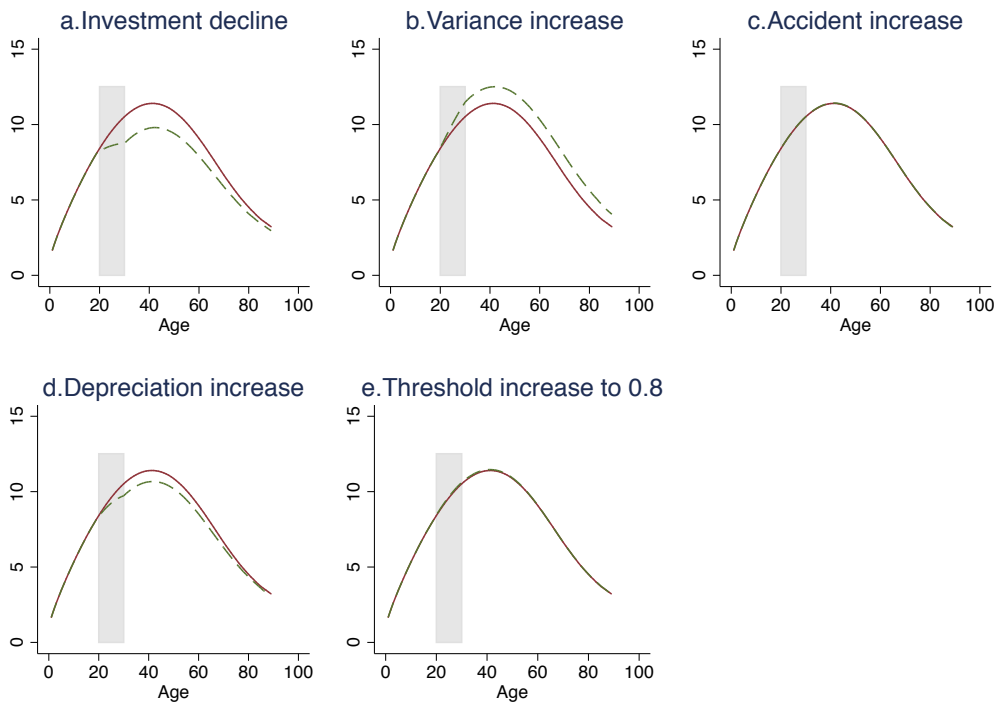
Note: Mortality rates are shown in log(10) scale. The estimated parameters are in Appendix Table 4.2.

Figure 4.16: US Age-specific Mortality rates per 1,000 in 1990, by age and cause of death



Note: this figure is reproduced from Schwandt and von Wachter (2018)'s paper "Mortality Profiles of Unlucky Cohorts: Effects of Entering the Labor Market in a Recession on Longevity" who generously agreed to let us use it. The data come from period (not cohort tables) so they are not directly comparable to ours but we use it to demonstrate that the mortality rate from non-disease related causes of death is well approximated by a step function turns on in adolescence. Mortality rates are shown in log(10) scale.

Figure 4.17: Health effects of temporary shocks at age 20

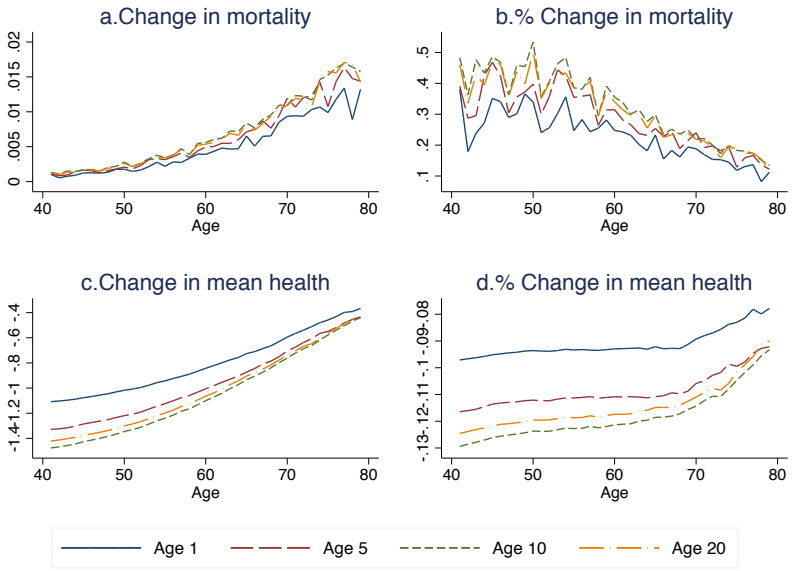


Year	Territorial Changes
1861	Annexion of *departements* of Savoie and Haute-Savoie, and of *Comte* de Nice
1869	Franco-Prussian war: loss of Alsace-Lorraine
1914-	WWI: East of France, from Nord Pas-de-Calais to Vosges, is occupied by German military.
1919	At the end of WWI, Alsace-Lorraine is re-integrated to French territory
1939	WW2: Loss of Alsace-Lorraine
1943	WW2: Loss of Corsica
1945	Current territory: Alsace-Lorraine and Corsica are re-integrated to French territory

These changes in territory results in large changes in the population and death counts. This is illustrated below for population. It is unclear how to compute mortality in the year of the change. We compute it by using a weighted average of the population at the beginning and end of the year..

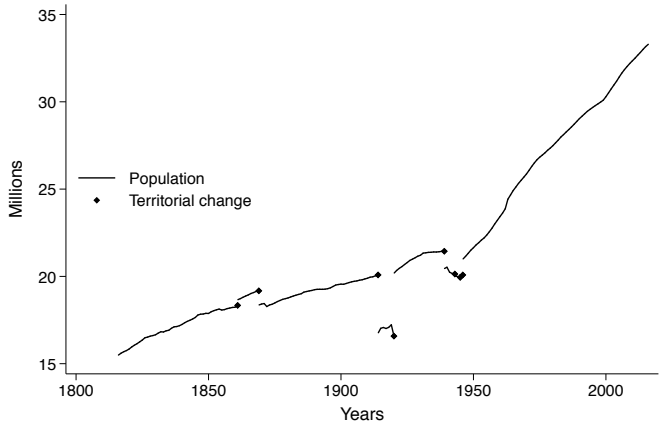
Migration. In the HMD cohort population counts are available. However, because of

Figure 4.18: Effects of a temporary decrease in investments in childhood on health and mortality, by age at the time of the shock



Note: Simulations for two populations of 500,000 individuals each. The shaded area corresponds to the years of the temporary shock, which is a 50% decline in I for 10 years. The baseline parameters are the same as in Figure 4.2b but we add an adolescent hump at age 14 with $\kappa = 0.001$. Mortality rates are shown in log(10) scale.

Figure 4.19: Female population in France since 1816



Note: See the technical documentation of the Human Mortality Database for details about the population coverage for the French mortality data.

migrations, these counts cannot be used to derive a survival curve for a cohort. Because of net positive immigration occurring in France, the number of individuals in a given cohort can even increase from one year to the next. This is especially true at the end of the Algerian War. (e.g. the size of the female cohort born in 1910 increases from $300,369$ to $303,273$ between 1962 and 1963, despite a reported mortality rate of 0.5162. . The unit of analysis in our model of mortality is a country cohort, hence abstracts from migration. In our model the mortality rates coincide exactly with the slope of the survival curve. This is not true in the HMD. The population of the cohort melts natives and immigrants of the same age.

2. Computing the death rates, survival rates and life expectancy

Death rates. When taking our model to the data we target the most direct counterpart of our modeled cohort "mortality rate", which is computed as the number of individuals who died during a year, divided by the number of individuals alive at the beginning of the day. In typical life tables this number corresponds to what demographer call q_t, the probability of dying in a given year, and is conceptually distinct to the mortality rate, denoted by m_t. The main difference lies in adjusting the denominator — the size of the population. As more individuals die during the year the population needs to be adjusted to estimate the size of the remaining population exposed to the risk of death. Because our baseline model does not take this adjustment into account, we compute a direct counterpart of our theoretical object. Therefore, we compute the raw death rate in year t for a given cohort , q_t , as follows:

$$q_t = \frac{D_t}{N_t}$$

where D_t is the death count for year t from the HMD cohort table and N_t is the population on January 1st of year t. The HMD makes adjustments to compute a probability that is corrected for the fact that the data do not tract the same individuals over time, so the probability of dying is not correctly computed for a given cohort. The q we estimate with the raw counts is very similar to what is reported by the HMD except for the first year of life and the last years of life as shown in Figure 4.6. This results in our under-estimating life expectancy somewhat.

Survival curves. We compute the survival curve recursively as follows. After initializing $S_0 = 100$, we iteratively compute:

$$S_t = S_{t-1} \times (1 - q_{t-1})$$

Life expectancy. Life Expectancy (LE) is an important statistics for the health profile of a given cohort. We compute LE as a way of comparing our model to the data in a parsimonious way. While we try to provide informative estimates of cohort life expectancy, we do not claim that their accuracy is comparable to demographic studies. Nevertheless, as we treat the series generated by our model in exactly the same manner as the data series, we obtain pairs of LE that are readily comparable.

4. Estimation routine

We compute our estimates using Matlab's canned fminsearch routine, a downhill simplex method, and Powell (1964)'s conjugate direction method. We first estimate the model using fminsearch until the objective function changes by less than 10^{-3} . We then use these estimates as starting values for Powell's routine. Once Powell's routine converges, we use the estimated values from this procedure and implement fminsearch again until it converges.

5. Bootstrapping standard errors

Estimates from sample data come with standard errors. However, the mortality rates in the HMD are computed from birth certificates of the total population, not a sample of it. A typical cohort in our study counts $400,000$ individuals. As a results, the standard errors are negligible and therefore do not report them for the French cohorts.

In contrast, we do compute the standard errors for the monkey estimates as the data in that case consist of samples of one or two hundreds of individuals. One way of bootstrapping the standard errors it, given a series of mortality rates for a cohort, to view each sample of size N as a sequence of Bernoulli trials with varying success rates. Alternatively, one can view the survival curve of a population of size N as an $N \times 1$ vector of age at death. One

can produce bootstrap estimates by drawing with replacement M subsamples of size S and compute the empirical survival curve.

Bibliography

Ackerberg, Daniel A, Kevin Caves, and Garth Frazer (2015). "Identification properties of recent production function estimators". In: *Econometrica* 83.6, pp. 2411–2451.

Aghion, Philippe et al. (2017). *Tax simplicity and heterogeneous learning*. Tech. rep. National Bureau of Economic Research.

Allain, Marie-Laure et al. (2017). "Retail mergers and food prices: Evidence from France". In: *The Journal of Industrial Economics* 65.3, pp. 469–509.

Almond, Douglas and Janet Currie (2011). "Killing me softly: The fetal origins hypothesis". In: *The Journal of Economic Perspectives* 25.3, pp. 153–172. URL: `http://www.ingentaconnect.com/content/aea/jep/2011/00000025/00000003/art00008` (visited on 02/24/2017).

Almond, Douglas, Janet Currie, and Valentina Duque (2017). *Childhood Circumstances and Adult Outcomes: Act II*. Tech. rep. National Bureau of Economic Research. URL: `http://www.nber.org/papers/w23017` (visited on 02/24/2017).

Andreoni, James, Brian Erard, and Jonathan Feinstein (1998). "Tax compliance". In: *Journal of economic literature* 36.2, pp. 818–860.

Antras, Pol (2004). "Is the US aggregate production function Cobb-Douglas? New estimates of the elasticity of substitution". In: *Contributions in Macroeconomics* 4.1.

Antràs, Pol et al. (2012). "Measuring the upstreamness of production and trade flows". In: *American Economic Review* 102.3, pp. 412–16.

Armitage, Peter and Richard Doll (1954). "The age distribution of cancer and a multi-stage theory of carcinogenesis". In: *British journal of cancer* 8.1, p. 1.

Arrow, Kenneth J et al. (1961). "Capital-labor substitution and economic efficiency". In: *The review of Economics and Statistics*, pp. 225–250.

Ashenfelter, Orley, Daniel Hosken, and Matthew Weinberg (2014). "Did Robert Bork understate the competitive impact of mergers? Evidence from consummated mergers". In: *The Journal of Law and Economics* 57.S3, S67–S100.

Atalay, Enghin (2017). "How important are sectoral shocks?" In: *American Economic Journal: Macroeconomics* 9.4, pp. 254–80.

Atkeson, Andrew and Ariel Burstein (2008). "Pricing-to-market, trade costs, and international relative prices". In: *American Economic Review* 98.5, pp. 1998–2031.

Atkeson, Andrew and Patrick J Kehoe (1999). "Models of energy use: Putty-putty versus putty-clay". In: *American Economic Review* 89.4, pp. 1028–1043.

Atkeson, Andrew and Patrick J. Kehoe (2005). "Modeling and Measuring Organization Capital". In: *Journal of Political Economy* 113.5, pp. 1026–1053. DOI: 10.1086/431289. eprint: https://doi.org/10.1086/431289. URL: https://doi.org/10.1086/431289.

Autor, David et al. (2017). *The fall of the labor share and the rise of superstar firms*. National Bureau of Economic Research.

Azar, José, Ioana Marinescu, and Marshall I Steinbaum (2017). *Labor market concentration*. Tech. rep. National Bureau of Economic Research.

Azar, José, Martin C Schmalz, and Isabel Tecu (2018). "Anticompetitive effects of common ownership". In: *The Journal of Finance* 73.4, pp. 1513–1565.

Azar, José and Xavier Vives (2018). "Oligopoly, macroeconomic performance, and competition policy". In:

Bachas, Pierre and Mauricio Soto (2018). *Not (ch) your average tax system: corporate taxation under weak enforcement*.

Bachmann, Rudiger, Ricardo J. Caballero, and Eduardo M. R. A. Engel (Oct. 2013). "Aggregate Implications of Lumpy Investment: New Evidence and a DSGE Model". In: *American Economic Journal: Macroeconomics* 5.4, pp. 29–67. DOI: 10.1257/mac.5.4.29. URL: http://www.aeaweb.org/articles?id=10.1257/mac.5.4.29.

Bagger, Jesper et al. (2014). "Tenure, experience, human capital, and wages: A tractable equilibrium search model of wage dynamics". In: *American Economic Review* 104.6, pp. 1551–96.

Bain, Joe Staten (1956). *Barriers to new competition: their character and consequences in manufacturing industries*. Vol. 329. Harvard University Press Cambridge, MA.

Baqaee, David Rezza and Emmanuel Farhi (2017). *Productivity and Misallocation in General Equilibrium.* Tech. rep. National Bureau of Economic Research.

— (2018). *The Microeconomic Foundations of Aggregate Production Functions*. Tech. rep.

Barrot, Jean-Noël and Julien Sauvagnat (2016). "Input specificity and the propagation of idiosyncratic shocks in production networks". In: *The Quarterly Journal of Economics* 131.3, pp. 1543–1592.

Basu, Susanto and John G Fernald (1997). "Returns to scale in US production: Estimates and implications". In: *Journal of political economy* 105.2, pp. 249–283.

Becker, Randy A et al. (2006). "Micro and macro data integration: The case of capital". In: *A new architecture for the US national accounts*. University of Chicago Press, pp. 541–610.

Beguin, Jean-Marc and Olivier Haag (2017). *Methodologie de la statistique annuelle d'entreprises, Description du systeme Esane.*

Benzarti, Youssef and Jarkko Harju (2018). "Are Taxes turning human into machines: Using Payroll variation to estimate capital-labor elasticity of substitution". In:

Besanko, David and Daniel F Spulber (1993). "Contested mergers and equilibrium antitrust policy". In: *The Journal of Law, Economics, and Organization* 9.1, pp. 1–29.

Blanchard, Olivier J, William D Nordhaus, and Edmund S Phelps (1997). "The medium run". In: *Brookings Papers on Economic Activity* 1997.2, pp. 89–158.

Blonigen, Bruce A and Justin R Pierce (2016). *Evidence for the effects of mergers on market power and efficiency*. Tech. rep. National Bureau of Economic Research.

Boehm, Johannes, Swati Dhingra, John Morrow, et al. (2016). "Swimming upstream: input-output linkages and the direction of product adoption". In: *CEP Discussion Paper* 1407.

Bos, Iwan and Joseph E Harrington Jr (2010). "Endogenous cartel formation with heterogeneous firms". In: *The RAND Journal of Economics* 41.1, pp. 92–117.

Bronikowski, Anne M. et al. (2011). "Aging in the natural world: comparative data reveal similar mortality patterns across primates". In: *Science* 331.6022, pp. 1325–1328. URL: `http://science.sciencemag.org/content/331/6022/1325.short` (visited on 02/24/2017).

Caballero, Ricardo J and Eduardo MRA Engel (1999). "Explaining investment dynamics in US manufacturing: a generalized (S, s) approach". In: *Econometrica* 67.4, pp. 783–826.

Caballero, Ricardo J and Mohamad L Hammour (1998). "Jobless growth: appropriability, factor substitution, and unemployment". In: *Carnegie-Rochester Conference Series on Public Policy*. Vol. 48. Elsevier, pp. 51–94.

Caliendo, Lorenzo, Ferdinando Monte, and Esteban Rossi-Hansberg (2015). "The anatomy of French production hierarchies". In: *Journal of Political Economy* 123.4, pp. 809–852.

Calvo, Guillermo A (1976). "Optimal growth in a putty-clay model". In: *Econometrica: Journal of the Econometric Society*, pp. 867–878.

Carrillo, Paul, Dina Pomeranz, and Monica Singhal (Apr. 2017). "Dodging the Taxman: Firm Misreporting and Limits to Tax Enforcement". In: *American Economic Journal: Applied Economics* 9.2, pp. 144–64. DOI: 10.1257/app.20140495. URL: http://www.aeaweb.org/articles?id=10.1257/app.20140495.

Case, Anne and Angus S Deaton (2005). "Broken down by work and sex: How our health declines". In: *Analyses in the Economics of Aging*. University of Chicago Press, pp. 185–212.

Case, Anne, Darren Lubotsky, and Christina Paxson (2002). "Economic status and health in childhood: The origins of the gradient". In: *The American Economic Review* 92.5, pp. 1308–1334. URL: http://www.ingentaconnect.com/content/aea/aer/2002/00000092/00000005/art00003 (visited on 02/28/2017).

Cass, David and Joseph E Stiglitz (1969). "The implications of alternative saving and expectations hypotheses for choices of technique and patterns of growth". In: *Journal of Political Economy* 77.4, Part 2, pp. 586–627.

Ceci-Renaud, Nila and Paul-Antoine Chevalier (2010). "L'impact des seuils de 10, 20 et 50 salaries sur la taille des entreprises francaises". In: *Economie et Statistique* 437.

Chamberlain, Gary (1986). "Asymptotic efficiency in semi-parametric models with censoring". In: *journal of Econometrics* 32.2, pp. 189–218.

Chaney, Thomas (2014). "The network structure of international trade". In: *American Economic Review* 104.11, pp. 3600–3634.

Chen, Zhao et al. (2017). "Notching R&D Investment with Corporate Income Tax Cuts in China". In:

Chetty, Raj (2009). "Is the taxable income elasticity sufficient to calculate deadweight loss? The implications of evasion and avoidance". In: *American Economic Journal: Economic Policy* 1.2, pp. 31–52.

— (2012). "Bounds on elasticities with optimization frictions: A synthesis of micro and macro evidence on labor supply". In: *Econometrica* 80.3, pp. 969–1018.

Chetty, Raj, John N Friedman, et al. (2011). "Adjustment costs, firm responses, and micro vs. macro labor supply elasticities: Evidence from Danish tax records". In: *The quarterly journal of economics* 126.2, pp. 749–804.

Chetty, Raj, Michael Stepner, et al. (2016). "The association between income and life expectancy in the United States, 2001-2014". In: *Jama* 315.16, pp. 1750–1766. URL: `http://jamanetwork.com/journals/jama/fullarticle/2513561?utm_term=alsomay` (visited on 03/02/2017).

Chirinko, Robert S (2008). "σ : *Thelongandshortofit*". In: *Journal of Macroeconomics* 30.2, pp. 671–686.

Collard-Wexler, Allan and Jan De Loecker (2016). *Production function estimation with measurement error in inputs*. Tech. rep. National Bureau of Economic Research.

Contoyannis, Paul, Andrew M. Jones, and Nigel Rice (2004). "The dynamics of health in the British Household Panel Survey". In: *Journal of Applied Econometrics* 19.4, pp. 473–503. URL: `http://onlinelibrary.wiley.com/doi/10.1002/jae.755/full` (visited on 03/18/2017).

Cooper, Russell W and John C Haltiwanger (2006). "On the nature of capital adjustment costs". In: *The Review of Economic Studies* 73.3, pp. 611–633.

Cooper, Russell, John Haltiwanger, and Jonathan L Willis (2007). "Search frictions: Matching aggregate and establishment observations". In: *Journal of Monetary Economics* 54, pp. 56–78.

Costa, Dora L (2012). "Scarring and mortality selection among Civil War POWs: A long-term mortality, morbidity, and socioeconomic follow-up". In: *Demography* 49.4, pp. 1185–1206.

Currie, Janet and Mark Stabile (2003). "Socioeconomic Status and Child Health: Why Is the Relationship Stronger for Older Children?" In: *The American Economic Review* 93.5, pp. 1813–1823. URL: http://www.jstor.org/stable/3132154 (visited on 03/02/2017).

Cutler, David M. et al. (2012). "The Oxford Handbook of Health Economics". In:

Dahl, Ronald E et al. (2018). "Importance of investing in adolescence from a developmental science perspective". In: *Nature* 554.7693, p. 441.

Dalgaard, Carl-Johan and Holger Strulik (2014). "Optimal aging and death: understanding the Preston curve". In: *Journal of the European Economic Association* 12.3, pp. 672–701.

De Loecker, Jan and Jan Eeckhout (2017). *The rise of market power and the macroeconomic implications*. Tech. rep. National Bureau of Economic Research.

Deaton, Angus S. and Christina Paxson (1994). "Saving, growth, and aging in Taiwan". In: *Studies in the Economics of Aging*. University of Chicago Press, pp. 331–362. URL: http://www.nber.org/chapters/c7349.pdf (visited on 03/18/2017).

Deaton, Angus S. and Christina H. Paxson (1997). "The effects of economic and population growth on national saving and inequality". In: *Demography* 34.1, pp. 97–114. URL: http://link.springer.com/article/10.2307/2061662 (visited on 03/18/2017).

— (1998). "Aging and inequality in income and health". In: *The American Economic Review* 88.2, pp. 248–253. URL: http://www.jstor.org/stable/116928 (visited on 03/18/2017).

Decker, Ryan A et al. (2017). "Declining dynamism, allocative efficiency, and the productivity slowdown". In: *American Economic Review* 107.5, pp. 322–26.

Deschênes, Olivier and Michael Greenstone (2011). "Climate change, mortality, and adaptation: Evidence from annual fluctuations in weather in the US". In: *American Economic Journal: Applied Economics* 3.4, pp. 152–85.

Deschenes, Olivier and Enrico Moretti (2009). "Extreme weather events, mortality, and migration". In: *The Review of Economics and Statistics* 91.4, pp. 659–681.

Di Giovanni, Julian, Andrei A Levchenko, and Isabelle Mejean (2014). "Firms, destinations, and aggregate fluctuations". In: *Econometrica* 82.4, pp. 1303–1340.

Diez, Mr Federico, Mr Daniel Leigh, and Suchanan Tambunlertchai (2018). *Global market power and its macroeconomic implications*. International Monetary Fund.

Disease Control, Centers for, Prevention, et al. (2014). "Summary health statistics: National health interview survey". In: *Atlanta, GA: Centers for Disease Control and Prevention*.

Doraszelski, Ulrich and Jordi Jaumandreu (2018). "Measuring the Bias of Technological Change". In: *Journal of Political Economy* 126.3, pp. 1027–1084. DOI: 10.1086/697204. URL: https://doi.org/10.1086/697204.

Duffie, Darrell, Gaston Giroux, and Gustavo Manso (2010). "Information percolation". In: *American Economic Journal: Microeconomics* 2.1, pp. 100–111.

Duffie, Darrell, Semyon Malamud, and Gustavo Manso (2009). "Information percolation with equilibrium search dynamics". In: *Econometrica* 77.5, pp. 1513–1574.

— (2014). "Information percolation in segmented markets". In: *Journal of Economic Theory* 153, pp. 1–32.

Edmond, Chris, Virgiliu Midrigan, and Daniel Yi Xu (2015). "Competition, markups, and the gains from international trade". In: *American Economic Review* 105.10, pp. 3183–3221.

— (2018). *How costly are markups?* Tech. rep. National Bureau of Economic Research.

Einav, Liran, Amy Finkelstein, and Paul Schrimpf (2015). "The response of drug expenditure to nonlinear contract design: evidence from medicare part D". In: *The quarterly journal of economics* 130.2, pp. 841–899.

Elsby, Michael WL, Bart Hobijn, and Ayşegül Şahin (2013). "The decline of the US labor share". In: *Brookings Papers on Economic Activity* 2013.2, pp. 1–63.

Fack, Gabrielle and Camille Landais (2016). "The effect of tax enforcement on tax elasticities: Evidence from charitable contributions in France". In: *Journal of Public Economics* 133, pp. 23–40.

Galama, Titus J and Hans Van Kippersluis (2018). "A Theory of Socio-economic Disparities in Health over the Life Cycle". In: *The Economic Journal* 129.617, pp. 338–374.

Garicano, Luis, Claire Lelarge, and John Van Reenen (Nov. 2016). "Firm Size Distortions and the Productivity Distribution: Evidence from France". In: *American Economic Review*

106.11, pp. 3439–79. DOI: 10.1257/aer.20130232. URL: http://www.aeaweb.org/articles?id=10.1257/aer.20130232.

Gattaz, Yvon (1979). *La fin des patrons*. Robert Laffont.

Gavrilov, Leonid Anatol?evich and Natal?i?a? Sergeevna Gavrilova (1991). "The biology of life span: a quantitative approach." In:

Gilchrist, Simon and John C Williams (2000). "Putty-clay and investment: a business cycle analysis". In: *Journal of Political Economy* 108.5, pp. 928–960.

Giovanni, Julian di, Andrei A Levchenko, and Isabelle Mejean (2018). "The micro origins of international business-cycle comovement". In: *American Economic Review* 108.1, pp. 82–108.

Gompertz, Benjamin (1820). "A Sketch of an Analysis and Notation Applicable to the Estimation of the Value of Life Contigencies". In: *Philosophical Transactions of the Royal Society of London* 110, pp. 214–332. URL: http://www.jstor.org/stable/107559 (visited on 02/23/2017).

— (1825). "On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies". In: *Philosophical transactions of the Royal Society of London* 115, pp. 513–583. URL: http://www.jstor.org/stable/107756 (visited on 02/23/2017).

— (1862). "A Supplement to Two Papers Published in the Transactions of the Royal Society," On the Science Connected with Human Mortality;" The One Published in 1820, and the Other in 1825". In: *Philosophical Transactions of the Royal Society of London* 152, pp. 511–559. URL: http://rstl.royalsocietypublishing.org/content/152/511.full.pdf (visited on 02/23/2017).

— (1871). "On one uniform law of mortality from birth to extreme old age, and on the law of sickness". In: *Journal of the Institute of Actuaries and Assurance Magazine* 16.5, pp. 329–344. URL: http://www.jstor.org/stable/41135309 (visited on 02/23/2017).

Gourio, Francois (2011). "Putty–clay technology and stock market volatility". In: *Journal of Monetary Economics* 58.2, pp. 117–131.

Gourio, François and Nicolas Roys (2014). "Size-dependent regulations, firm size distribution, and reallocation". In: *Quantitative Economics* 5.2, pp. 377–416.

Grossman, Michael (1972). "On the concept of health capital and the demand for health". In: *Journal of Political economy* 80.2, pp. 223–255. URL: http://www.journals.uchicago.edu/doi/pdfplus/10.1086/259880 (visited on 02/23/2017).

Grullon, Gustavo, Yelena Larkin, and Roni Michaely (2015). "The disappearance of public firms and the changing nature of US industries". In: 2612047.

Gutiérrez, Germán and Thomas Philippon (2017). "Investmentless Growth: An Empirical Investigation". In: *Brookings Papers on Economic Activity* 2017.2, pp. 89–190.

Guvenen, Fatih et al. (2017). *Offshore profit shifting and domestic productivity measurement*. Tech. rep. National Bureau of Economic Research.

Harrington Jr, Joseph E and Yanhao Wei (2017). "What can the duration of discovered cartels tell us about the duration of all cartels?" In: *The Economic Journal* 127.604, pp. 1977–2005.

Havari, Enkelejda and Franco Peracchi (2017). "Growing up in wartime: Evidence from the era of two world wars". In: *Economics & Human Biology* 25, pp. 9–32.

Havez, Pierre (2018). "Les seuils sociaux en France, Quel impact sur la démographie des PME?" In: *Master Thesis, Paris-I Sorbonne*.

Heckman, James (1990). "Varieties of selection bias". In: *The American Economic Review* 80.2, p. 313.

Heligman, Larry and John H. Pollard (1980). "The age pattern of mortality". In: *Journal of the Institute of Actuaries* 107.01, pp. 49–80. URL: http://journals.cambridge.org/abstract_S0020268100040257 (visited on 02/24/2017).

Hicks, John (1932). *The theory of wages*.

Hopenhayn, Hugo A (2014). "Firms, misallocation, and aggregate productivity: A review". In: *Annual Review of Economics* 6.1, pp. 735–770.

Hottman, Colin J, Stephen J Redding, and David E Weinstein (2016). "Quantifying the sources of firm heterogeneity". In: *The Quarterly Journal of Economics* 131.3, pp. 1291–1364.

House, James S, Paula M Lantz, and Pamela Herd (2005). "Continuity and change in the social stratification of aging and health over the life course: evidence from a nationally representative longitudinal study from 1986 to 2001/2002 (Americans' Changing Lives Study)". In: *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences* 60.Special_Issue_2, S15–S26.

Houthakker, Hendrik S (1955). "The Pareto distribution and the Cobb-Douglas production function in activity analysis". In: *The Review of Economic Studies* 23.1, pp. 27–31.

Hsieh, Chang-Tai and Peter J Klenow (2009). "Misallocation and manufacturing TFP in China and India". In: *The Quarterly journal of economics* 124.4, pp. 1403–1448.

Hsieh, Chang-Tai and Benjamin A Olken (2014). "The Missing" Missing Middle"". In: *Journal of Economic Perspectives* 28.3, pp. 89–108.

Hugonnier, Julien, Benjamin Lester, and Pierre-Olivier Weill (2018). *Frictional intermediation in over-the-counter markets*. Tech. rep. National Bureau of Economic Research.

Hummer, Robert A. and Joseph T. Lariscy (2011). "Educational attainment and adult mortality". In: *International handbook of adult mortality*. Springer, pp. 241–261. URL: http://link.springer.com/chapter/10.1007/978-90-481-9996-9_12 (visited on 03/14/2017).

Hurst, Erik, Geng Li, and Benjamin Pugsley (2014). "Are household surveys like tax forms? Evidence from income underreporting of the self-employed". In: *Review of economics and statistics* 96.1, pp. 19–33.

Ibragimov, I. (1956). "On the Composition of Unimodal Distributions". In: *Theory of Probability & Its Applications* 1.2, pp. 255–260. DOI: 10.1137/1101021.

Jäger, Kirsten (2017). "EU KLEMS Growth and Productivity Accounts 2017 Release, Statistical Module1". In: *The Conference Board*.

Johansen, Leif (1959). "Substitution versus fixed production coefficients in the theory of economic growth: a synthesis". In: *Econometrica: Journal of the Econometric Society*, pp. 157–176.

Kalecki, Michał (1945). "On the Gibrat distribution". In: *Econometrica: Journal of the Econometric Society*, pp. 161–170.

Karabarbounis, Loukas and Brent Neiman (2014). "The global decline of the labor share". In: *The Quarterly Journal of Economics* 129.1, pp. 61–103.

Kesternich, Iris et al. (2014). "The effects of World War II on economic and health outcomes across Europe". In: *Review of Economics and Statistics* 96.1, pp. 103–118.

Khan, Aubhik and Julia K Thomas (2008). "Idiosyncratic shocks and the role of nonconvexities in plant and aggregate investment dynamics". In: *Econometrica* 76.2, pp. 395–436.

Kim, Kyoo, Amil Petrin, and Suyong Song (2016). "Estimating production functions with control functions when capital is measured with error". In: *Journal of Econometrics* 190.2, pp. 267–279.

Klette, Tor Jakob and Samuel Kortum (2004). "Innovating firms and aggregate innovation". In: *Journal of political economy* 112.5, pp. 986–1018.

Kleven, Henrik J and Mazhar Waseem (2013). "Using notches to uncover optimization frictions and structural elasticities: Theory and evidence from Pakistan". In: *The Quarterly Journal of Economics* 128.2, pp. 669–723.

Kleven, Henrik Jacobsen (2016). "Bunching". In: *Annual Review of Economics* 8.1, pp. 435–464. DOI: 10.1146/annurev-economics-080315-015234. URL: https://doi.org/10.1146/annurev-economics-080315-015234.

Krusell, Per et al. (2000). "Capital-skill complementarity and inequality: A macroeconomic analysis". In: *Econometrica* 68.5, pp. 1029–1053.

La Rochebrochard, Elise de (2000). "Age at puberty of girls and boys in France: Measurements from a survey on adolescent sexuality". In: *Population: An English Selection*, pp. 51–79. URL: http://www.jstor.org/stable/3030244 (visited on 05/03/2017).

Lando, David (2004). "Credit Risk Modeling: Theory and Applications". In:

Lee, Chulhee (2017). "Long-term health consequences of prenatal exposure to the Korean War". In: *Asian Population Studies* 13.1, pp. 101–117.

León-Ledesma, Miguel A, Peter McAdam, and Alpo Willman (2010). "Identifying the elasticity of substitution with biased technical change". In: *American Economic Review* 100.4, pp. 1330–57.

Levenstein, Margaret C and Valerie Y Suslow (2006). "What determines cartel success?" In: *Journal of economic literature* 44.1, pp. 43–95.

Levhari, David (1968). "A note on Houthakker's aggregate production function in a multifirm industry". In: *Econometrica: journal of the Econometric Society*, pp. 151–154.

Levinsohn, James and Amil Petrin (2003). "Estimating production functions using inputs to control for unobservables". In: *The Review of Economic Studies* 70.2, pp. 317–341.

Li, Ting and James Anderson (2013). "Shaping human mortality patterns through intrinsic and extrinsic vitality processes". In: *Demographic research* 28, pp. 341–372. URL: `http://www.demographic-research.org/volumes/vol28/12/default.htm` (visited on 02/24/2017).

Limpert, Eckhard, Werner A. Stahel, and Markus Abbt (2001). "Log-normal distributions across the sciences: Keys and clues on the charms of statistics, and how mechanical models resembling gambling machines offer a link to a handy way to characterize log-normal distributions, which can provide deeper insight into variability and probability—normal or log-normal: That is the question". In: *BioScience* 51.5, pp. 341–352. URL: `http://bioscience.oxfordjournals.org/content/51/5/341.short` (visited on 03/02/2017).

Liu, Li and Ben Lockwood (2015). "VAT notches". In:

Liu, Zuyun et al. (2019). "The Role of Epigenetic Aging in Education and Racial/Ethnic Mortality Disparities Among Older US Women". In: *Psychoneuroendocrinology*.

Loudon, Irvine (2000). "Maternal mortality in the past and its relevance to developing countries today". In: *The American journal of clinical nutrition* 72.1, 241s–246s. URL: `http://ajcn.nutrition.org/content/72/1/241s.short` (visited on 02/23/2017).

Lucas Jr, Robert E and Benjamin Moll (2014). "Knowledge growth and the allocation of time". In: *Journal of Political Economy* 122.1, pp. 1–51.

Lucas, Robert E (1969). "Labor-capital substitution in US manufacturing". In: *The taxation of income from capital*, pp. 223–274.

— (1978). "On the size distribution of business firms". In: *The Bell Journal of Economics*, pp. 508–523.

Luttmer, Erzo GJ (2010). "Models of growth and firm heterogeneity". In: *Annual Review of Economics* 2.1, pp. 547–576.

Maksimovic, Vojislav, Gordon Phillips, and Liu Yang (2013). "Private and public merger waves". In: *The Journal of Finance* 68.5, pp. 2177–2217.

Marschak, Jacob and William H Andrews (1944). "Random simultaneous equations and the theory of production". In: *Econometrica, Journal of the Econometric Society*, pp. 143–205.

Marx, Benjamin M (2015). "Dynamic bunching estimation and the cost of reporting regulations for charities". In: *MPRA Paper #88647*.

Miyagiwa, Kaz and Chris Papageorgiou (2007). "Endogenous aggregate elasticity of substitution". In: *Journal of Economic Dynamics and Control* 31.9, pp. 2899–2919.

Neary, J. Peter (2003). "Globalization and Market Structure". In: *Journal of the European Economic Association* 1, pp. 245–271. DOI: 10.1162/154247603322390928. URL: https://onlinelibrary.wiley.com/doi/abs/10.1162/154247603322390928.

Neven, Damien J and Lars-Hendrik Röller (2005). "Consumer surplus vs. welfare standard in a political economy model of merger control". In: *International Journal of Industrial Organization* 23.9-10, pp. 829–848.

Nocke, Volker and Michael D Whinston (2010). "Dynamic merger review". In: *Journal of Political Economy* 118.6, pp. 1200–1251.

— (2013). "Merger policy with merger choice". In: *American Economic Review* 103.2, pp. 1006–33.

O'brien, Daniel P and Steven C Salop (1999). "Competitive effects of partial ownership: Financial interest and corporate control". In: *Antitrust LJ* 67, p. 559.

Oberfield, Ezra and Devesh Raval (2014). *Micro data and macro technology*. Tech. rep. National Bureau of Economic Research.

Occhino, Filippo, Kim Oosterlinck, and Eugene N. White (2006). *How Occupied France Financed Its Own Exploitation in World War II*. Tech. rep. National Bureau of Economic Research. URL: http://www.nber.org/papers/w12137 (visited on 03/21/2017).

Olley, G Steven and Ariel Pakes (1992). *The dynamics of productivity in the telecommunications equipment industry*. Tech. rep. National Bureau of Economic Research.

Ottaviani, Marco and Abraham L Wickelgren (2011). "Ex ante or ex post competition policy? A progress report". In: *International Journal of Industrial Organization* 29.3, pp. 356–359.

Palloni, Alberto and Hiram Beltrán-Sánchez (2016). "Demographic Consequences of Barker Frailty". In: *Dynamic Demographic Analysis*. Springer, pp. 147–176. URL: http://link.springer.com/chapter/10.1007/978-3-319-26603-9_8 (visited on 02/24/2017).

Perla, Jesse and Christopher Tonetti (2014). "Equilibrium imitation and growth". In: *Journal of Political Economy* 122.1, pp. 52–76.

Phelps, Edmund S (1963). "Substitution, fixed proportions, growth and distribution". In: *International Economic Review* 4.3, pp. 265–288.

Pompei, Francesco and Richard Wilson (2002). "A quantitative model of cellular senescence influence on cancer and longevity". In: *Toxicology and industrial health* 18.8, pp. 365–376.

Preston, Samuel, Patrick Heuveline, and Michel Guillot (2000). "Demography: measuring and modeling population processes". In: URL: http://www.citeulike.org/group/18479/article/12819791 (visited on 02/23/2017).

Remund, Adrien, Carlo G. Camarda, and Tim Riffe (June 2018). "A Cause-of-Death Decomposition of Young Adult Excess Mortality". In: *Demography* 55.3, pp. 957–978. ISSN: 1533-7790. DOI: 10.1007/s13524-018-0680-9. URL: https://doi.org/10.1007/s13524-018-0680-9.

Restuccia, Diego and Richard Rogerson (2008). "Policy distortions and aggregate productivity with heterogeneous establishments". In: *Review of Economic dynamics* 11.4, pp. 707–720.

Ricardo, David (1817). *On the principles of political economy and taxation*. London: John Murray.

Saez, Emmanuel (Aug. 2010). "Do Taxpayers Bunch at Kink Points?" In: *American Economic Journal: Economic Policy* 2.3, pp. 180–212. DOI: 10.1257/pol.2.3.180. URL: http://www.aeaweb.org/articles?id=10.1257/pol.2.3.180.

Salop, Steven C (2017). "Invigorating vertical merger enforcement". In: *Yale LJ* 127, p. 1962.

Sargent, Thomas J (1978). "Estimation of dynamic labor demand schedules under rational expectations". In: *Journal of Political Economy* 86.6, pp. 1009–1044.

Satō, Kazuo (1975). *Production functions and aggregation*. Vol. 90. North-Holland.

Schiman, Jeffrey C, Robert Kaestner, and Anthony T Lo Sasso (2017). *Early Childhood Health Shocks and Adult Wellbeing: Evidence from Wartime Britain*. Tech. rep. National Bureau of Economic Research.

Scholz, John T and Neil Pinney (1995). "Duty, fear, and tax compliance: The heuristic basis of citizenship behavior". In: *American Journal of Political Science*, pp. 490–512.

Schwartz, Joel (2000). "Harvesting and long term exposure effects in the relation between air pollution and mortality". In: *American journal of epidemiology* 151.5, pp. 440–448.

Selten, Reinhard (1973). "A simple model of imperfect competition, where 4 are few and 6 are many". In: *International Journal of Game Theory* 2.1, pp. 141–201.

Sharrow, David J. and James J. Anderson (2016). "Quantifying Intrinsic and Extrinsic Contributions to Human Longevity: Application of a Two-Process Vitality Model to the Human Mortality Database". In: *Demography* 53.6, pp. 2105–2119. URL: `http://link.springer.com/article/10.1007/s13524-016-0524-4` (visited on 02/24/2017).

Smagghue, Gabriel (2014). "Size-Dependent Regulation and Factor Income Distribution". In: *Unpublished manuscript*.

Solow, Robert M (1957). "Technical change and the aggregate production function". In: *The review of Economics and Statistics*, pp. 312–320.

— (1962). "Substitution and fixed proportions in the theory of capital". In: *The Review of Economic Studies* 29.3, pp. 207–218.

Sorkin, Isaac (2015). "Are there long-run effects of the minimum wage?" In: *Review of economic dynamics* 18.2, pp. 306–333.

Stigler, George J (1964). "A theory of oligopoly". In: *Journal of political Economy* 72.1, pp. 44–61.

Stuart, G (1945). "Diphtheria incidence in European countries". In: *British medical journal* 2.4426, p. 613.

Symeonidis, George (2008). "The effect of competition on wages and productivity: evidence from the United Kingdom". In: *The Review of Economics and Statistics* 90.1, pp. 134–146.

Thiele, Thorvald Nicolai (1871). "On a mathematical formula to express the rate of mortality throughout the whole of life, tested by a series of observations made use of by the Danish Life Insurance Company of 1871". In: *Journal of the Institute of Actuaries* 16.5, pp. 313–329.

Tirole, Jean (1988). *The theory of industrial organization*. MIT press.

Vaupel, James W., Kenneth G. Manton, and Eric Stallard (1979). "The impact of heterogeneity in individual frailty on the dynamics of mortality". In: *Demography* 16.3, pp. 439–454. URL: http://link.springer.com/article/10.2307/2061224 (visited on 02/23/2017).

Weibull, Waloddi (1951). "Wide applicability". In: *Journal of applied mechanics* 103.730, pp. 293–297.

Wilcox, Allen J. and Ian T Russell (1983). "Birthweight and perinatal mortality: I. On the frequency distribution of birthweight". In: *International Journal of Epidemiology* 12.3, pp. 314–318. URL: http://ije.oxfordjournals.org/content/12/3/314.short (visited on 02/24/2017).

Williamson, Oliver E (1968). "Economies as an antitrust defense: The welfare tradeoffs". In: *The American Economic Review* 58.1, pp. 18–36.

Wong, Chloe Chung Yi et al. (2010). "A longitudinal study of epigenetic variation in twins". In: *Epigenetics* 5.6, pp. 516–526.

Zeger, Scott L, Francesca Dominici, and Jonathan Samet (1999). "Harvesting-resistant estimates of air pollution effects on mortality". In: *Epidemiology*, pp. 171–175.