HARVARD UNIVERSITY
Graduate School of Arts and Sciences

DISSERTATION ACCEPTANCE CERTIFICATE

The undersigned, appointed by the

Department of Economics

have examined a dissertation entitled

"Essays on the Consequences of Innovative Industries"

presented by **Benjamin Niswonger**

candidate for the degree of Doctor of Philosophy and hereby certify that it is worthy of acceptance.

Signature _____

Typed name: Prof. Elhanan Helpman

Signature _____

Typed name: Prof. Edward Glaeser

Signature _____

Typed name: Prof. Oliver Hart

Date: February 8, 2023

Essays on the Consequences of Innovative Industries

A dissertation presented

by

Benjamin Niswonger

to

The Department of Economics

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Economics

Harvard University

Cambridge, Massachusetts

February 2023

Dissertation Advisors:
Professor Elhanan Helpman
Professor Edward Glaeser

Author:
Benjamin Niswonger

Essays on the Consequences of Innovative Industries

Abstract

This dissertation develops novel modeling techniques, theoretical insights and empirical estimations in order to uncover a diverse set of downstream consequences from innovative industries.

Chapter 1, which is co-authored with Elhanan Helpman, provides a theoretical framework which allows us to analyze the dynamics of industries constituted by a continuum of small firms and a finite number of large multi-product firms. The presence of a competitive fringe of small firms allows us to provide analytical results which are consistent with observed patterns of increased markups and concentration and a decreased labor share. Furthermore, the model predicts an inverted-U relationship between product span and productivity with implications for industry dynamics.

Chapter 2 focuses on the response to changes in skill demand precipitated by the proliferation of skill-biased technologies. The main focus is on the spatial winners and losers from rising skill premiums. This chapter highlights the impact that skill acquisition which is biased in favor of local-skill demand can have on spatial inequality and economy-wide productivity. The main takeaway is that there is large scope for a trade-off between static productivity gains from agglomeration and dynamic gains from local signaling.

Chapter 3, co-authored with Zoë Hitzig, considers how the rise of new industries might affect the optimal regulation of bilateral contracts. We begin by building a mechanism design framework which theoretically motivates the use of default and immutable laws in the regulation of contracts. We consider a setting with observable but unverifiable information

which is common knowledge to two agents which efficiently bargain over their joint surplus. Our mechanism design framework makes clear how the principal optimally uses defaults in order to achieve equity objectives and sets immutable laws to internalize externalities. We then apply this framework to the current debate on the classification of gig-workers in order to highlight the tradeoffs between these two forms of regulation when the regulator cannot achieve first best.

Contents

Title Page i
Copyright ii
Abstract iii
Table of Contents v
List of Tables vii
List of Figures viii
Acknowledgments x

Introduction **1**

1 Dynamics of Markups, Concentration and Product Span **4**
1.1 Introduction 4
1.2 Preliminaries 7
1.3 Entry of Single-Product Firms 10
1.4 Transition Dynamics 12
1.5 Comparative Dynamics 20
1.5.1 Costs of Single-Product Firms 27
1.6 Optimal Allocation 30
1.7 Conclusion 35

2 What You See is What You Get: Local Labor Markets and Skill Acquisition **37**
2.1 Introduction 37
2.2 Suggestive Evidence 41
2.2.1 Complexity in College Major Choice for Potential Income 41
2.2.2 College Major Choice is Consistent with Local Degree Distribution . 42
2.2.3 Inter-County Migration has Slowed 45
2.3 Structural Spatial Model 47
2.3.1 Preferences: 47
2.3.2 Learning: 48
2.3.3 Signal Extraction Model: 49
2.3.4 Expected Indirect Utility: 50

2.3.5 Goods Production . . . 51
2.3.6 Housing . . . 52
2.3.7 Solving the Model . . . 52
2.3.8 Comparing to the Existing Literature . . . 54
2.4 Model Simplification and Estimation . . . 55
2.4.1 Data . . . 55
2.4.2 Non-Tradable Component and Amenities . . . 55
2.4.3 Signaling Impact . . . 57
2.4.4 Agglomeration Elasticity . . . 60
2.5 Experiment . . . 60
2.6 Conclusion . . . 63

3 Bilateral Contracts and Social Welfare **65**
3.1 Introduction . . . 65
3.2 Examples: Incomplete Contracts and Social Welfare . . . 71
3.3 Model . . . 78
3.4 First Best Analysis . . . 81
3.4.1 Implementation of efficient contracts . . . 81
3.4.2 Default delegation with equity concerns . . . 82
3.4.3 Default delegation with externality concerns . . . 87
3.4.4 Limitations of default delegation . . . 89
3.4.5 Interpreting results . . . 92
3.5 Second Best Analysis: An Application to the Regulation of Platform Work . 95
3.5.1 Set up . . . 95
3.5.2 Firm control . . . 98
3.5.3 Equal bargaining . . . 105
3.5.4 Worker control . . . 109
3.6 Conclusion . . . 110

References **112**

Appendix A Appendix to Chapter 1 **118**
A.1 Comparative Dynamics . . . 118
A.2 Empirical Analysis . . . 119
A.3 Optimal Allocation . . . 121
A.4 Comparative Statics: Given Number of Brands . . . 126
A.4.1 Aggregative Economy . . . 129

Appendix B Appendix to Chapter 2 **137**
B.1 Why Focus on Learning 137
B.1.1 Major Switching Between Freshman and Senior Year 137
B.1.2 Information Flows Through Local Major and Industry Composition . 138

Appendix C Appendix to Chapter 3 **141**
C.1 Numerical Example: The Uniform Commercial Code and “Reasonable" Defaults 141
C.2 Proofs 144
C.3 Max-Min Social Welfare Functions 150
C.4 Extensions 152
C.4.1 Inequity penalty in SWF. 152
C.4.2 Exogenous income 153
C.5 Application Derivations and Discussion 154
C.5.1 Regulator Problem 155
C.5.2 Equal Bargaining $\delta = .5$ 159
C.5.3 Firm Power $\delta = 0$ 161
C.5.4 Worker Power $\delta = 1$ 165

List of Tables

2.1 Effect of Local Labor Market on Major Choice 45
2.2 High and Low Amenity MSAs 57
2.3 Estimates of Key Model Parameters Based on Migration Probabilities 59
2.4 Impact of Local College Fraction in Dollar Terms 60
2.5 Estimating Agglomeration Effect of Destination College Fraction 60
2.6 Impact of Destination College Fraction on Wage 61

3.1 Direct mechanism for implementing (q_θ, c_θ) in state θ. 84
3.2 Direct mechanism with message-independent threats ($|\Theta| = 2$). 85
3.3 Direct mechanism with message-independent threats ($|\Theta| = 3$). 90
3.4 Summary of results: First-best implementation with default delegation . . . 92
3.5 Examples of Defaults and Limits in Settings from Section 3.2 92

A.1 Average Number of Product Lines vs. Productivity Deciles 119
A.2 Quadratic Relationship of Productivity on Product Span 120

B.1 Effect of Labor Market Similarities on Major Switching 138
B.2 Migration Based on Labor Market Similarities 139

C.1 Direct mechanism for implementing (q_θ, c_θ) in state θ. 145
C.2 Default delegation with max-min regulator ($|\Theta| = 3$). 151

List of Figures

1.1 Transition Dynamics 16
1.2 Dynamics of the market share in response to a decline in the marginal cost a_i 25
1.3 Average Number of Product Lines vs. Labor Productivity Deciles 26
1.4 Number of Segments vs. Labor Productivity 27
1.5 Dynamics of market shares in response to a decline in P 29

2.1 ACS Total Income by Major for Prime Aged College Graduates 43
2.2 Comparing Working Population Major Composition vs. College Student Stated Major 44
2.3 CPS Intercounty Migration Rates 46
2.4 Impact of Redistributing College Workers on Welfare 62

3.1 (a) Bargaining from an Explicit Prespecified Contract; (b) Bargaining from the U.C.C.'s "Reasonable" Default 73
3.2 An Unforeseen Work Regime Requiring a New Default 76
3.3 Distribution of Surplus Across States with Varying Default Quality 100
3.4 Relationship Between Externality and Equity on Optimal Default Delegation 103

A.1 Number of Industries vs. Labor Productivity 120

C.1 (a) Bargaining from an Explicit Prespecified Contract (q_0, c_0); (b) Bargaining from the U.C.C.'s "Reasonable" Default $(q_d(\omega), c_d(\omega))$ 142
C.2 Numerical Example: (a) Explicit Prespecified Contract (q_0, c_0); (b) U.C.C.'s "Reasonable" Default $(q_d(\omega), c_d(\omega))$ 143

Acknowledgments

I would like to thank Elhanan Helpman, Ed Glaeser and Oliver Hart each of whom had a major influence on the chapters of this dissertation. I would especially like to thank Elhanan who bore the brunt of helping me stay focused and make progress on my research and for guiding me more generally through the the research process. I also appreciate my conversations with Ed which stimulated many new ideas and directions. Oliver's interest and excitement for the idea behind the third chapter of this dissertation was a consistent source of motivation. I would also like to thank the many other professors who gave me advice and feedback throughout the PhD including Eric Maskin, Shengwu Li, David Yang, Marc Melitz, Pol Antras, Steve Marglin and Jerry Green.

I would also like to thank the many friends I have made at the department who have helped iron out many of my projects and, more importantly, kept my spirits high. Zoë Hitzig, Dan Ramos-Menchelli, John Macke and Ria Granzier-Nakajima were tremendous coauthors at different points throughout the program and provided invaluable friendships. Similarly, Ariel Gomez, Ambra Seck, José Ramón Enríquez, Ambi Brunnel, Audrey Tiew, Anthony Yu, Robbie Minton, Giorgio Saponaro, Jennifer Zou, Ricardo Rodriguez-Padilla, Ashesh Rambachan and Myles Wagner are all to thank for hanging out, having fun and talking econ with me.

Lastly, I would like to thank my family, friends and especially my wife for providing reassurance and comfort throughout the program. I also want to mention the great states of New Hampshire, Vermont, Massachusetts, Maine, Kentucky, Texas, Utah and New Mexico for being lovely hosts at different points in the program as well as Guadaloupe, New Orleans, Las Vegas, West Palm Beach and Buenos Aires.

To Fernanda Lavalle and Mom and Dad

Introduction

This dissertation provides novel modeling frameworks and insights to the study of innovative industries. The three chapters below focus on industry, labor market and regulatory dynamics, respectively. Overall, the dissertation is a response to several macroeconomic trends that have been the focus of considerable research effort: increased markups and concentration, changing labor demand, and declining labor shares (Autor *et al.* (2020a); De Loecker *et al.* (2020), Autor (2019), Autor *et al.* (2020a)). These pivotal features of the economy are multifaceted and interact in different ways depending on what economic actors are in focus. In this dissertation, I consider three distinct lenses on these key issues.

In the first chapter, coauthored with Elhanan Helpman, we develop a modeling framework which allows us to consider industry dynamics when firms are able to invest in increasing their product span through R&D. In order to capture trends in market concentration and markups, we consider investment decisions made by oligopolistic firms. These settings typically lead to differential games which don't allow for simple characterizations of the solution. To make this setting tractable, we incorporate the assumption that there exists a competitive fringe of firms which make entry decisions after large firms make their investment decisions. This assumption is an extreme version of the stylized fact that small firms exhibit greater turnover than the establishments of large firms.

With this assumption, we are able to solve for the optimal investment path for large firms analytically. The results are consistent with the evidence on changes in market concentration, labor share and markups. Beyond that our model provides novel insights into the cross-sectional evolution of firms in an industry. Most notably, our model predicts

an inverted-U relationship between firm labor productivity and product span. Furthermore, the response to specific shocks depend on the labor productivity of the firm. For instance, an increase in the technology of firms in the competitive fringe will lead to dynamic paths where lower productivity firms will contract their product span over time whereas high productivity firms will expand. We provide empirical evidence for this inverted-U property using Compustat data.

In order to focus on industry dynamics, we made simplifying assumptions with regards to labor and skill demand. However, there is a vast literature which highlights how recent technological advancements have been skill-biased and have replaced routine cognitive and non-cognitive tasks (Autor (2019)). The second chapter of this dissertation considers how the labor market responds to changes in skill demand. The primary focus is on the spatial distribution of skills which has been an important feature in the debate on place-based policies. I add to this literature by considering the role that local labor markets play on skill *acquisition*.

Students tend to be biased towards majors which are in relatively high demand in their local labor market. This can have significant consequences as the return to different majors vary drastically. There is positive externality due to the presence of workers with high return majors in a student's local labor market. This implies that although there may be agglomerative benefits of concentrating skilled workers in a small number of labor markets, this may come at the expense of the total supply of skilled workers. I highlight this potential by developing a structural spatial model which incorporates educational and locational choice. In an empirical exercise, I show how a policy which reduces the concentration of in-demand skill could reduce spatial inequality and increase overall productivity.

In response to the decreasing labor share and worker bargaining power brought about by changes in skill demand, the government has recently been considering the regulation of large innovative firms. In the third chapter of this dissertation, my coauthor, Zoë Hitzig, and I provide theoretical guidance for the regulation of bilateral contracts between firms and workers when the government has efficiency, equity and externality concerns. We focus

on the regulation of gig-work which has led to considerable debate and various, at times contradictory, policies. Using a mechanism design framework we build off of the literature on contract law to highlight the use of default and immutable rules. Specifically, we show that with observable but unverifiable information that is common between the contracting parties and a binary state space, the principal (i.e. the regulator) can obtain the first best outcome by setting immutable rules which internalize externalities and default rules which achieve the optimal distribution of surplus. We go on to show the tradeoffs between default and immutable rules in an applied setting where the state space is continuous.

Overall, this dissertation aims to provide modeling frameworks which allow us to study optimal policies in the context of innovative industries from three distinct vantage points: firms, workers and the interaction between the two. Each chapter aims to show how the changes in underlying technologies can have implications for firm, worker and worker-firm dynamics which necessitate policies which are adaptive: the optimal policy varies with industry structure, labor demand and worker-firm relationship. These chapters should be thought of as an entryway to considering these distinct and important features of our current innovation economy.

Chapter 1

Dynamics of Markups, Concentration and Product Span[1]

1.1 Introduction

A number of recent studies have investigated the evolution of markups and the growth of concentration in U.S. industries, finding that both markups and concentration have increased; see Autor *et al.* (2020a); De Loecker *et al.* (2020). These studies find that the ascent in average markups was driven by rising markups of the largest firms and market share reallocation from low- to high-markup firms. Contemporaneously, the labor share declined. We develop a model of firm dynamics that generates these patterns, as well as rich predictions about the unfolding of the cross-section of firms. Our theory focuses on the evolution of a sector rather than on long-run growth of the entire economy.[2]

An industry has a continuum of varieties of a differentiated product and it is populated by a continuum of single-product firms and a finite number of large multi-product firms.

[1]Co-authored with Elhanan Helpman. Copyright American Economic Association; reproduced with permission.

[2]While the model in the main text does not exhibit long-run growth, we describe in the Appendix a model with a continuum of sectors that is suitable for long-run growth analysis. In that model every sector is similar to the sector analyzed in the main text of the paper.

While the turnover of single-product firms is very high, the large multi-product firms have long life spans. Large firms lose some products over time, but they can invest in innovation in order to replenish or expand the range of their products. Free entry of the single-product firms, which engage in monopolistic competition, creates a competitive fringe that impacts the oligopolistic competition of the large firms. The interaction between the single- and multi-product firms plays a key role in the dynamics of this industry, both during transition and in steady state.[3]

Our assumptions capture salient features of the data. According to Cao *et al.* (2019), 95% of firms in the U.S. economy are single-establishment firms, but their share in employment is only 45%. Furthermore, Kehrig and Vincent (2019) report that during 1972-2007 an average of 72% of the plants in manufacturing belonged to single-plant firms and 28% belonged to multi-plant firms (see their Table A1). At the same time, single-plant firms manufactured 22% of value added compared to 78% of the value added manufactured by multi-plant firms. Both studies point out that firm growth took place mostly through the extensive margin, by opening new plants that often produced new products. In addition, Hsieh and Rossi-Hansberg (2020) provide evidence that firm growth through the acquisition of new product lines played an important role in the business strategies of U.S. corporations in three major sectors: services, retail and wholesale. Growth of firms in these sectors was driven by expansions to new locations, i.e., new product lines. Finally, studying the size distribution of firms between 1995 and 2014, Cao *et al.* (2019) conclude that the largest contributors to the increase in the number of establishments per firm were declining costs of external innovation and declining exit rates.

The theory developed in this paper predicts that large multi-product firms grow through innovation that expands their product lines, eventually reaching a steady state. In the

[3]Interactions between a monopolistically competitive fringe of single-product firms and oligopolistic large firms have been studied in a static framework by Shimomura and Thisse (2012) in a closed economy and by Parenti (2018) in an open economy. Cavenaile *et al.* (2019) develop a model of endogenous growth with quality-ladders, in which there is a fringe of competitive small firms that produce a homogenous good. In addition, there are single-product large firms that produce different varieties of a product. Their model is mostly quantitative, used to study the relationship between innovation and competition.

processes, these firms raise their markups and reduce their labor share. As a result, the average markup—measured with either cost-share or sales-share weights—rises and the aggregate labor share declines. Both the cost-weighted and the sales-weighted markups rise due to rising individual markups of multi-product firms and the reallocation of market shares from single- to multi-product firms.

The steady state size distribution of firms is driven by heterogeneity of labor productivity, with more productive firms having larger market shares. Nevertheless, this monotonic relationship does not translate into a monotonic relationship between productivity and product span. The reason is that the marginal profitability of investment in innovation is larger when manufacturing costs are lower, but larger market shares reduce the incentives to invent new product lines. This tension can produce an inverted-U relationship between labor productivity and product span in the cross-section of firms. Using the Compustat data for 2018, we provide evidence in support this prediction. We also show that the inverted-U relationship between labor productivity and product span stems from distortions in the market equilibrium. In the optimal dynamic allocation product span is an increasing function of labor productivity in the steady state.

Our model predicts that improvements in the technology of single-product firms, which raise the competitive pressure on the multi-product oligopolies, lead to a decline in the market share of every large firm on impact. Still, the resulting transition dynamics to a new steady state vary across the large firms according to size. In particular, multi-product firms with large market shares compensate for the initial loss of competitiveness (reflected in the loss of market share) by gradually expanding their product span and raising their market share over time, while firms with small market shares further reduce their product span and market shares over time. As a result, the size distribution of multi-product firms becomes more unequal in the new steady state.

We describe some basic elements of the model in the next section. In Section 1.3 we detail the entry decisions of single-product firms and their impact on the pricing strategy of large firms. These results are then used in Section 1.4 to study the innovation decisions of

large firms and the resulting transition dynamics. We show that whenever multi-product firms widen their product span in the transition, they grow in size and so do their markups, while the labor share declines. In the following section, Section 1.5, we study comparative dynamics, some of which were described above. In Section 1.6 we characterize the optimal dynamic allocation and describe a set of policy measures that implement the optimal allocation. Section 1.7 concludes.

1.2 Preliminaries

We consider an economy with a continuum of individuals of mass 1. The labor market is competitive and every individual earns the same wage rate.

There are two sectors. One sector produces a homogeneous good with one unit of labor per unit output and there is always positive demand for its product. We normalize the price of this good to equal one. Therefore the competitive wage also equals one. The other sector produces varieties of a differentiated product.[4]

Every individual supplies a fixed amount of labor, l, and has a utility function:[5]

$$u = x_0 + \frac{\varepsilon}{\varepsilon - 1}\left[\int_0^N x(\omega)^{\frac{\sigma-1}{\sigma}} d\omega\right]^{\frac{(\varepsilon-1)\sigma}{\varepsilon(\sigma-1)}}, \ \sigma > \varepsilon > 1, \tag{1.1}$$

where x_0 is consumption of the homogeneous good, $x(\omega)$ is consumption of variety ω of the differentiated product, σ is the elasticity of substitution between varieties of the differentiated product and ε gauges the degree of substitutability between varieties of the differentiated product and the homogeneous good. The assumption $\sigma > \varepsilon$ asserts that brands of the differentiated product are better substitutes for each other than for the homogeneous good. The assumption $\varepsilon > 1$ ensures that aggregate spending on the differentiated product declines when its price rises (see below).

[4]It is straightforward to generalize the analysis to multiple sectors with differentiated products.

[5]We can add to this utility function a disutility of effort $\psi \frac{l^{1+\nu}}{1+\nu}$, $\nu, \psi > 0$, as is common in some of the macro literature. Due to the quasi-linearity of the utility function, this would lead every individual to optimally choose a constant supply of labor, $l = \psi^{-1/\nu}$. For this reason we simplify by assuming that l is constant.

Real consumption of the differentiated product is:

$$X = \left[\int_0^N x(\omega)^{\frac{\sigma-1}{\sigma}} d\omega \right]^{\frac{\sigma}{\sigma-1}}. \tag{1.2}$$

Using this definition, the price index of X is:

$$P = \left[\int_0^N p(\omega)^{1-\sigma} d\omega \right]^{\frac{1}{\sigma-1}},$$

where $p(\omega)$ is the price of variety ω. In this setup an individual chooses consumption to maximize utility subject to the budget constraint $x_0 + PX = l + y$, where y is non-wage income. This yields $X = P^{-\varepsilon}$ as long as consumers purchase the homogenous good and varieties of the differentiated product, which we assume always to be the case (this requires l to be large enough). Clearly, in this case the demand for variety ω is:

$$x(\omega) = P^{\delta} p(\omega)^{-\sigma}, \delta = \sigma - \varepsilon > 0. \tag{1.3}$$

Aggregate spending on the differentiated product equals $PX = P^{1-\varepsilon}$, which declines in P, because $\varepsilon > 1$. An individual's consumption choice yields the indirect utility function

$$V = l + y + \frac{1}{\varepsilon - 1} P^{1-\varepsilon}, \tag{1.4}$$

where the third term on the right-hand side represents consumer surplus.

Two types of firms operate in sector X: atomless single-product firms and m large multi-product firms. Every large firm has a positive measure of product lines (recall the discussion in the introduction of evidence in support of this specification). Single-product firms produce a total of $\bar{r} > 0$ varieties, each one specializing in a single brand. Large firm i has $r_i > 0$ product lines, $i = 1, 2, ..., m$. All the brands supplied to the market are distinct from each other.

All single-product firms share the same technology, which requires $\bar{a}$ unit of labor per unit output.[6] Facing the demand function (1.3), a single-product firm maximizes profits

[6]It is straightforward to allow for heterogeneity of the single-product firms, by assuming that each one

$P^{\delta} p(\omega)^{-\sigma} [p(\omega) - \bar{a}]$, taking as given the price index P. Therefore, a single-product firm prices its brand ω according to $p(\omega) = \overline{p}$, where:

$$\overline{p} = \frac{\sigma}{\sigma - 1} \bar{a}. \tag{1.5}$$

This yields the standard markup $\overline{\mu} = \sigma / (\sigma - 1)$ for a monopolistically competitive firm.

A large firm i has a technology that requires a_i units of labor per unit output, and it faces the demand function (1.3) for each one of its brands. As a result, it prices every brand equally. We denote this price by p_i. The firm chooses p_i to maximize profits $r_i P^{\delta} p_i^{-\sigma} (p_i - a_i)$. However, unlike a single-product firm, a large firm does not view P as given, because it recognizes that

$$P = \left(\overline{r}\, \overline{p}^{1-\sigma} + \sum_{j=1}^{m} r_j p_j^{1-\sigma} \right)^{\frac{1}{1-\sigma}}, \tag{1.6}$$

and therefore that its pricing policy has a measurable impact on the price index of the differentiated product. Accounting for this dependence of P on the firm's price, the profit maximizing price is:

$$p_i = \frac{\sigma - \delta s_i}{\sigma - \delta s_i - 1} a_i, \tag{1.7}$$

where s_i is the market share of firm i and:[7]

$$s_i = \frac{r_i p_i^{1-\sigma}}{P^{1-\sigma}} = \frac{r_i p_i^{1-\sigma}}{\overline{r}\, \overline{p}^{1-\sigma} + \sum_{j=1}^{m} r_j p_j^{1-\sigma}}. \tag{1.8}$$

Equations (1.7) and (1.8) jointly determine prices and market shares of large firms. The markup factor of firm i is $\mu_i = (\sigma - \delta s_i)/(\sigma - \delta s_i - 1)$, which is increasing in its market share. When the market share equals zero the markup is $\sigma/(\sigma - 1)$, the same as the markup of a single product firm. The markup factor varies across firms as a result of differences in either the product span, r_i, or the marginal production cost, a_i. We analyze the dependence

of them draws a unit labor requirement from a known distribution. Since this type of heterogeneity plays no essential role in our analysis, we have chosen to work with the simpler formulation.

[7] Note that $\sigma - \delta s_i - 1 = \sigma (1 - s_i) + \varepsilon s_i - 1 > 0$ and $\sigma - \delta s_i = \sigma (1 - s_i) + \varepsilon s_i > 0$.

of prices, market shares and markups on marginal costs and product spans in the next section.

1.3 Entry of Single-Product Firms

The number of large firms is given and we analyze in the next section the evolution of their product spans, r_i. Unlike large firms, single-product firms enter the industry until their profits equal zero. Firms in this sector play a two-stage game: in the first stage single-product firms enter; in the second stage all firms play a Bertrand game as described in the previous section. Under these circumstances, (1.5) and (1.7) portray the equilibrium prices, except that the number of single product firms, $\bar{r}$, is endogenous. We seek to characterize a subgame perfect equilibrium of this stage game.

To determine the equilibrium number of single-product firms, assume that they face an entry cost f and they enter until profits equal zero. In a subgame perfect equilibrium every entrant correctly forecasts the number of entrants, and the price that will be charged for every variety in the second stage of the game. Therefore, every single-product firm correctly forecasts the price index P. Using the optimal price (1.5) and the profit function $P^{\delta}\overline{p}^{-\sigma}\left(\overline{p}-\bar{a}\right)$, this free entry condition can be expressed as:

$$\frac{1}{\sigma}P^{\delta}\left(\frac{\sigma}{\sigma-1}\bar{a}\right)^{1-\sigma}=f. \tag{1.9}$$

The left-hand side of this equation describes the operating profits, which equal a fraction $1/\sigma$ of revenue, while the right-hand side represents the entry cost. In these circumstances the price index P is determined by f and $\bar{a}$, and it is rising in both f and $\bar{a}$. Importantly, it does not depend on the number of large firms nor on their product spans.

We now use (1.7) and (1.8) to calculate the response of prices and market shares to changes in the number of product lines, changes in marginal costs and changes in the price index P. Denoting by a hat the proportional rate of change of a variable, i.e., $\hat{x}=dx/x$, differentiating these two equations yields the solutions:

$$\hat{p}_i = \frac{\beta_i}{1+(\sigma-1)\beta_i}\hat{r}_i + \frac{1}{1+(\sigma-1)\beta_i}\hat{a}_i + \frac{(\sigma-1)\beta_i}{1+(\sigma-1)\beta_i}\hat{P}, \tag{1.10}$$

$$\hat{s}_i = \frac{1}{1+(\sigma-1)\beta_i}\hat{r}_i - \frac{\sigma-1}{1+(\sigma-1)\beta_i}\hat{a}_i + \frac{\sigma-1}{1+(\sigma-1)\beta_i}\hat{P}, \tag{1.11}$$

where:

$$\beta_i = \frac{\delta s_i}{(\sigma-\delta s_i-1)(\sigma-\delta s_i)} > 0. \tag{1.12}$$

Due to the fact that the price index P responds neither to changes in r_i nor changes in a_i, an increase in r_i raises p_i and s_i, but it has no impact on prices and market shares of the other large firms. For the same reason, an increase in a_i raises p_i and reduces s_i, but has no impact on prices and market shares of the other large firms. Moreover, an increase in a_i raises the price of firm i less than proportionately, and therefore there is only partial pass-through of marginal costs to prices. The extent of the pass-through is smaller for a firm with a larger β_i, which is a firm with a larger market share. Finally, an increase in the price index P, which represents a decline in the competitive pressure in the industry, raises the price and the market share of *every* large firm. However, the price rises proportionately more and the market share rises proportionately less in firms with larger β_is, which are firms with larger market shares. Finally, the market share of a firm is larger the larger is its product span or the lower is its marginal cost of production. Noting again that the markup of every firm i is larger the larger its market share, we summarize these findings in

Proposition 1. *Suppose that the number of large firms and their product spans are given, but there is free entry of single-product firms. Then: (i) an increase in r_i raises the price, markup and market share of firm i, but has no impact on prices, markups and market shares of the other large firms; (ii) a decline in a_i reduces the price and raises the markup and market share of firm i, but has no impact on prices, markups and market shares of other large firms; (iii) a decline in the price index P, either due to a decline in $\bar{a}$ or a decline in f, reduces the price, markup and market share of every large firm,*

with prices changing proportionately more and market shares changing proportionately less for firms with initially larger market shares.

It is clear from this proposition that free entry of single-product firms leads large firms to compete for market share with single-product firms rather than with each other.[8] An increase in r_i or a decline in a_i, each of which raises the market share of firm i, does not impact the market share of other large firms, but do reduce the market share of single-product firms. Since the price index P does not change in response to changes in the number of product lines r_i or the marginal cost a_i, the market share of single-product firms must decline, which materializes through a decline in their joint product span. We, therefore, have

Proposition 2. *Suppose that the number of large firms and their product spans are given, but there is free entry of single-product firms. Then a decline in a_i or an increase in r_i reduces the number of single-product firms and their market share.*

1.4 Transition Dynamics

We next study the dynamics that arise when large firms can expand their product lines. Time is continuous and the economy starts at time $t = 0$. The range of products of firm i at time t is $r_i(t)$ for $t \geq 0$ and $r_i(0) = r_i^0$ is given.

Similarly to Klette and Kortum (2004), at every point in time firm i can invest to increase the number of its product lines. An investment flow of ι_i per unit time expands r_i by $\phi(\iota_i)$ units per unit time, where $\phi(\cdot)$ is the innovation function. We assume that $\phi(\iota)$ is increasing, concave, $\phi(0) = 0$ and it satisfies the Inada conditions $lim_{\iota \to 0} \phi'(\iota) = +\infty$ and $lim_{\iota \to +\infty} \phi'(\iota) = 0$. Furthermore, r_i depreciates at the rate θ per unit time, which randomly hits the continuum of available brands. It follows that the product span r_i satisfies the

[8] This is different, for example, from Atkeson and Burstein (2008), who have a continuum of industries, each one populated by a finite number of large single-product firms, and no small firms. Nevertheless, our and their pricing formulas have common elements.

differential equation:

$$\dot{r}_i = \phi(\iota_i) - \theta r_i, \text{ for all } t \geq 0, \tag{1.13}$$

where we have suppressed the time index t in $\dot{r}_i$, r_i and ι_i.

The endogenous expansion of product lines plays an important role in our theory. Hsieh and Rossi-Hansberg (2020) provide evidence that it also played an important role in the business strategies of U.S. corporations in three key sectors—services, retail and wholesale—where firm growth was dominated by expansion to new locations, i.e., new product lines. Cao *et al.* (2019) make a similar argument more broadly; firms grew predominantly on the extensive margin, through new establishments that often represented new product lines.[9]

At every point in time firms play a two stage game. In the first stage single-product firms enter and large firms invest in innovation. Single-product firms live only one instant of time. For this reason they make profits only in this single instant. This assumption represents an extreme form of the empirical property that the turnover of small firm establishments is much larger than the turnover of establishments of large firms. In the second stage all firms choose prices, in the manner described in the previous section. Under the circumstances the price index P is determined by the free entry condition (1.9), and it remains constant as long as the cost of entry and the cost of production of the single-product firms do no change.

In this economy the state vector is $\mathbf{r} = (r_1, r_2, ..., r_m)$, a function of time t, and the price p_i is a function of $\mathbf{r}$. Note, however, from (1.10) and (1.11) that p_i and s_i depend only on element r_i of $\mathbf{r}$. It follows that the profit flow of large firm i is:

$$\pi_i(\iota_i, r_i) = r_i P^{\delta} p_i(r_i)^{-\sigma} [p_i(r_i) - a_i] - \iota_i, \text{ for all } t \geq 0, \tag{1.14}$$

[9] Aghion *et al.* (2019) develop a model of economic growth in which the total number (measure) of product lines is constant, but a single firm can operate multiple product lines. They focus on explaining the decline in the long-run growth rate. The key trigger of their dynamics is a decline in a static cost function $c(n)$ that describes a firm's overhead cost of operating n product lines. They argue that a fall in these costs was caused by the IT revolution. Their firms have constant unit costs and the quality of products can be improved by investing in innovation, as in standard models of endogenous growth with quality ladders (see Grossman and Helpman (1991) and Aghion and Howitt (1992)). A firm acquires new product lines by gaining leadership positions through quality competition. They characterize a steady state of an economy with two types of firms—high- and low-productivity (unit labor requirements)—and study the impact of a decline in $c(n)$ on concentration, labor shares, the reallocation of market shares and the long-run growth rate.

where P is the same at every t and $p_i(r_i)$ is the price of firm i's brands as a function of r_i, given by (1.7). Evidently, if the firm's product span changes over time so do π_i, r_i, p_i and ι_i. Moreover, the firm's market share is also a function of r_i, $s_i(r_i)$. From (1.10) and (1.11) we obtain the elasticities of the function $p_i(r_i)$:

$$\frac{\partial p_i}{\partial r_i}\frac{r_i}{p_i} = \frac{\beta_i}{1+(\sigma-1)\beta_i}, \tag{1.15}$$

where β_i is defined in (1.12). Note that β_i is increasing in s_i and that, due to (1.11), s_i is increasing in r_i. Therefore β_i is increasing in r_i. As a result, the elasticity of the price function is larger the larger is s_i.

Next assume that the interest rate is constant and equal to ρ. This interest rate can be derived from the assumption that individuals discount future utility flows (1.1) with a constant rate ρ, so that they maximize the discounted present value of utility $\int_0^\infty e^{-\rho t}u(t)dt$. Under these circumstances firm i maximizes the discounted present value of its profits net of investment costs, π_i. It therefore solves the following optimal control problem:

$$\max_{\{\iota_i(t),r_i(t)\}_{t\geq 0}} \int_0^\infty e^{-\rho t}\pi_i\left[\iota_i(t), r_i(t)\right]dt$$

subject to (1.13), (1.14), $r_i(0) = r_i^0$, and a transversality condition to be described below, where $\pi_i(\iota_i, r_i)$ is defined in (1.14). In this problem ι_i is a control variable while r_i is a state variable. The current-value Hamiltonian is:

$$\mathcal{H}(\iota_i, r_i, \lambda_i) = \left\{r_i P^\delta p_i(r_i)^{-\sigma}\left[p_i(r_i) - a_i\right] - \iota_i\right\} + \lambda_i\left[\phi(\iota_i) - \theta r_i\right],$$

where λ_i is the co-state variable of constraint (1.13). The first-order conditions of this optimal control problem are:

$$\frac{\partial \mathcal{H}}{\partial \iota_i} = -1 + \lambda_i \phi'(\iota_i) = 0,$$

$$-\frac{\partial \mathcal{H}}{\partial r_i} = -\frac{\partial\left\{r_i P^\delta p_i(r_i)^{-\sigma}\left[p_i(r_i) - a_i\right]\right\}}{\partial r_i} + \theta\lambda_i = \dot{\lambda}_i - \rho\lambda_i,$$

and the transversality condition is:

$$\lim_{t\to\infty} e^{-\rho t}\lambda_i(t)\, r_i(t) = 0.$$

In addition, the optimal path of (ι_i, r_i) has to satisfy the differential equation (1.13).

The above first-order conditions can be expressed as:

$$\lambda_i \phi'(\iota_i) = 1, \tag{1.16}$$

$$\dot{\lambda}_i = (\rho+\theta)\lambda_i - P^{\delta} p_i(r_i)^{-\sigma}\left\{p_i(r_i) - a_i - r_i\left(\sigma p_i(r_i)^{-1}\left[p_i(r_i)-a_i\right]-1\right)p_i'(r_i)\right\}. \tag{1.17}$$

From (1.16) we obtain the investment level ι_i as an increasing function of λ_i, which we represent as $\iota_i(\lambda_i)$. Substituting this function into (1.13) yields the autonomous differential equation:

$$\dot{r}_i = \phi\left[\iota_i(\lambda_i)\right] - \theta r_i. \tag{1.18}$$

Next we substitute (1.7), (1.12) and (1.15) into (1.17) to obtain a second autonomous differential equation:

$$\dot{\lambda}_i = (\rho+\theta)\lambda_i - \Gamma_i(r_i), \tag{1.19}$$

where

$$\Gamma_i(r_i) = a_i^{1-\sigma} P^{\delta}\sigma\left[\frac{\sigma-\delta s_i(r_i)}{\sigma-\delta s_i(r_i)-1}\right]^{-\sigma}\frac{1}{\left[\sigma-\delta s_i(r_i)-1\right]\sigma + s_i(r_i)^2\delta^2} \tag{1.20}$$

represents the profitability of a new product line, given the firm's product span r_i; that is, it represents the *marginal* profitability of r_i. We show in the Appendix that this marginal profitability declines in r_i, i.e., $\Gamma_i'(r_i) < 0$.[10]

A solution to the autonomous system of differential equations (1.18) and (1.19) that satisfies the transversality condition is also a solution to the firm's optimal control problem, because $\mathcal{H}(\iota_i, r_i, \lambda_i)$ is concave in the first two arguments. This can be seen by observing

[10] By definition:

$$\Gamma_i(r_i) \equiv \frac{\partial\left\{r_i P^{\delta} p_i(r_i)^{-\sigma}\left[p_i(r_i)-a_i\right]\right\}}{\partial r_i},$$

where the right-hand side represents marginal profits of r_i. We show in the Appendix that $\Gamma_i(r_i)$ can be expressed as $\tilde{\Gamma}_i[s_i(r_i)]$, where $\tilde{\Gamma}_i(s_i)$ is a declining function (see (1.20) above and equation (A.1) in the Appendix). Since (1.11) implies that $s_i(r_i)$ is an increasing function, it follows that $\Gamma_i(r_i)$ is a declining function.

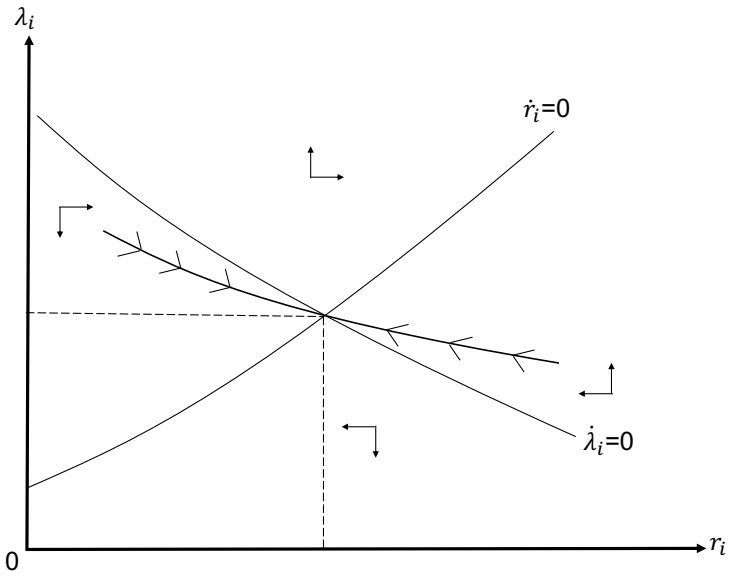


Figure 1.1: *Transition Dynamics*

that the Hamiltonian is additively separable in ι_i and r_i, and it is strictly concave in ι_i and in r_i. The steady state of these differential equations is characterized by:

$$\phi\left[\iota_i\left(\lambda_i\right)\right] = \theta r_i, \tag{1.21}$$

$$\left(\rho + \theta\right)\lambda_i = \Gamma_i\left(r_i\right). \tag{1.22}$$

The left-hand side of (1.21) is an increasing function of λ_i. Therefore, the curve in (r_i, λ_i) space along which r_i is constant is upward sloping. The right-hand side of (1.22) is declining in r_i, because $\Gamma_i'\left(r_i\right) < 0$. Therefore, the curve in (r_i, λ_i) space along which λ_i is constant is downward sloping. These curves are depicted in Figure 1.1. Based on the differential equations (1.18)-(1.19), the figure also depicts the resulting dynamics. There is a single stable saddle-path along which (r_i, λ_i) converge to the steady state and the transversality condition is satisfied in this steady state. On this saddle path either r_i rises and λ_i declines or r_i declines and λ_i rises, depending on whether r_i^0 is below or above its steady-state value.

Now suppose that all the r_i^0s are below their steady state values (this case arises, for example, when the economy is in steady state and innovation costs decline; see below).

Then every large firm expands its range of products over time. As a result, the number of single-product firms shrinks. This process continues until the economy reaches a steady state.

If at some point in time the number of single-product firms drops to zero, the dynamics change.[11] We focus, however, on the case in which $\bar{r} > 0$ for all $t \geq 0$. In this case the price index P remains constant as long as f and $\bar{a}$ do not change.

What can be said about the dynamics of profits net of investment costs? Changes of these profits over time can be expressed as (see (1.14)):

$$\begin{aligned}\frac{\partial \pi_i\left(\iota_i, r_i\right)}{\partial t} &= -\frac{\partial \iota_i}{\partial t} + \frac{\partial\left\{r_i P^{\delta} p_i\left(r_i\right)^{-\sigma}\left[p_i\left(r_i\right) - a_i\right]\right\}}{\partial r_i}\frac{\partial r_i}{\partial t} \\ &= -\frac{\partial \iota_i}{\partial t} + \Gamma_i\left(r_i\right)\frac{\partial r_i}{\partial t}.\end{aligned}$$

From (1.16) we see that ι_i is an increasing function of λ_i and λ_i declines in a firm that expands its product range. As a result, the firm's investment level ι_i declines over time, raising profits net of investment costs through a decline in investment spending. Moreover, $\Gamma_i\left(r_i\right) > 0$, and therefore an increase in r_i raises operating profits, thereby raising profits net of investment costs.[12] It follows that every firm that adds new product lines enjoys rising profits net of investment costs. Since wages are constant, this implies that the share of labor

[11] From that point on the optimal strategy of large firm i depends on the entire state vector $\mathbf{r}$. As a result, the firms engage in a differential game. Since no firm can commit to the entire path of its investments ι_i, one needs to adopt the closed loop solution to this game, in which the investment level ι_i is a function of the state vector $\mathbf{r}$. There do not exist user-friendly characterizations of solutions to such games. Instead, we provide in the Appendix an analysis of the impact of changes in the state variables r_i on prices, markups and market shares of the large firms.

[12] Intuitively, an increase in a product line raises directly a firm's profits due to the fact that the price exceeds marginal costs. But it also reduces profits indirectly as a result of an increase in the price in response to the expansion of product span. The former effect is represented by $p_i - a_i > 0$ in the tilted brackets of (1.17) while the latter effect is represented by $-r_i\left[\sigma p_i^{-1}\left(p_i - a_i\right) - 1\right] p_i' < 0$ in these brackets, noting that $p_i' > 0$ and $-\left[\sigma p_i^{-1}\left(p_i - a_i\right) - 1\right] = 1 - \sigma\left(1 - \mu_i^{-1}\right) < 0$. The last inequality results from the fact that the markup, $\mu_i = p_i / a_i$, is larger than $\sigma/\left(\sigma - 1\right)$ (see (1.7)). Nevertheless, the joint impact is positive, as shown in (1.20).

in national income declines when all large firms grow.[13]

Next consider the average markup in the differentiated product sector, defined as aggregate revenue divided by aggregate variable costs:

$$\mu_{av} = \frac{\overline{r}\,\overline{p}^{1-\sigma} + \sum_{j=1}^{m} r_j p_j^{1-\sigma}}{\overline{r}\,\overline{a}\,\overline{p}^{-\sigma} + \sum_{j=1}^{m} r_j a_j p_j^{-\sigma}}.$$

This statistic can be expressed as

$$\mu_{av} = \left(1 - \sum_{i=1}^{m} \varrho_i\right) \bar{\mu} + \sum_{i=1}^{m} \varrho_i \mu_i,$$

where

$$\varrho_i = \frac{r_i a_i p_i^{-\sigma}}{\overline{r}\,\overline{a}\,\overline{p}^{-\sigma} + \sum_{j=1}^{m} r_j a_j p_j^{-\sigma}} = \mu_{av} s_i \mu_i^{-1} \tag{1.23}$$

is the variable cost share of large firm i, $\bar{\mu} = \bar{p}/\bar{a}$ is the markup of a small firm and $\mu_i = p_i/a_i$ is the markup of large firm i. In other words, the average markup μ_{av} is a cost-weighted average of the markups of single- and multi-product firms. Next note from (1.23) that $\sum_{i=1}^{m} \varrho_i \mu_i = \mu_{av} \sum_{i=1}^{m} s_i$, and therefore:

$$\mu_{av} = \left(1 - \mu_{av} \sum_{i=1}^{m} s_i \mu_i^{-1}\right) \bar{\mu} + \mu_{av} \sum_{i=1}^{m} s_i,$$

which implies:

$$\mu_{av} = \frac{1}{\left(1 - \sum_{i=1}^{m} s_i\right) \bar{\mu}^{-1} + \sum_{i=1}^{m} s_i \mu_i^{-1}}. \tag{1.24}$$

When all large firms grow on the dynamic path, the market share of every one of them rises and so does its markup. The hikes in each firm's market share and markup consequently contribute to a rise in the average markup, μ_{av}, because the large firms have larger markups than the single-product firms.

While the cost-weighted average markup represents the ratio of aggregate revenue to

[13] In this economy aggregate income equals wages plus profits net of investment costs, i.e.,

$$y_{ag} = l + \sum_{j=1}^{m} \pi_i\left(\iota_i, r_i\right),$$

and the labor share is l/y_{ag}. When all large multi-product firms have r_i^0s below their steady state values they raise their product span over time and aggregate profits increase. As a result, the aggregate labor share declines.

aggregate variable cost, an alternative measure of average markups is a sales-share weighted average of the markups of all single- and multi-product firms (see, for example, Edmond *et al.* (2019)). In our model this average is:

$$\mu_{av}^{s} = \left(1 - \sum_{i=1}^{m} s_i\right) \bar{\mu} + \sum_{i=1}^{m} s_i \mu_i.$$

Since the markup of every large firm is higher than the markup of every single-product firm and the market share of every multi-product firm rises over time, this average markup increases over time. The growth in this average markup is driven by the same two forces that drive the rise in the cost-weighted average markup μ_{av}: rising markups of the large firms and market share reallocation from low-markup (single-product) to high-markup (multi-product) firms. We conclude that both measures of the average markup, μ_{av} and μ_{av}^{s}, are rising over time when large firms expand their product span.

Next compare the size of these markup statistics. Their ratio is given by:

$$\frac{\mu_{av}^{s}}{\mu_{av}} = \left[\left(1 - \sum_{i=1}^{m} s_i\right) \bar{\mu} + \sum_{i=1}^{m} s_i \mu_i\right] \left[\left(1 - \sum_{i=1}^{m} s_i\right) \bar{\mu}^{-1} + \sum_{i=1}^{m} s_i \mu_i^{-1}\right].$$

Since $1/\mu$ is a convex function of μ, Jensen's inequality implies that the right-hand side of this equation is larger than one, and therefore that $\mu_{av}^{s} > \mu_{av}$, which is what Edmond *et al.* (2019) found in the Compustat data.[14] We summarize these findings in

Proposition 3. *Consider an economy in which the initial range of products r_i^0 is smaller than its steady state value for every i, and in which $\bar{r} > 0$ at all times. Then over time: (i) every large firm i widens its product span, raises its markup, and experiences rising profits net of investment costs; (ii) the cost-weighted average markup and the sales-weighted average markup rise over time; (iii) the sales-average markup exceeds the cost-average markup at every point in time; and (iv) the share of*

[14]Using a dynamic model of monopolistic competition with a Kimball aggregator, Edmond *et al.* (2019) decompose the welfare cost of markups into three sources of influence: (i) aggregate markup; (ii) misallocation of inputs; and (iii) inefficiently low entry of firms. Their quantitative model implies that (i) accounts for 3/4 of the welfare cost while (ii) accounts for 1/4. The impact of entry is negligible. They also show that in the Compustat data the sales-weighted aggregate markup is higher than the cost-weighted aggregate markup, in line with our theoretical prediction, and that the gap between them has widened over time (see their Figure 8). Moreover, in their model the cost-weighted aggregate markup turns out to be the relevant measure for (i).

labor in national income declines over time.

Since wages are constant, so is wage income. Nonetheless, in view of Proposition 3(i), aggregate income—which consists of labor income plus aggregate profits net of investment costs—rises during the transition to a steady state. In view of the indirect utility function (1.4), this implies that aggregate utility rises over time (recall that P remains constant). Moreover, if this economy is populated by some individuals who own shares in large firms and other individuals who do not, the growth of large multi-product firms widens the disparity of well-being between these two groups.

1.5 Comparative Dynamics

For an economy that is initially in steady state, we study in this section the dynamics that arise in response to changes in the cost of inventing new product lines, the marginal costs of production and the cost of entry of single-product firms.

First, consider a change in the cost of innovation, as reflected in a shift of the function $\phi(\iota_i)$. We take $\kappa > 0$ to be a productivity measure of innovation and express the modified innovation function as $\kappa\phi(\iota_i)$. Initially $\kappa = 1$. An upward shift in κ represents a rise in the productivity of investment in innovation or a decline in innovation costs, while a decline in κ represents a decline in the productivity of investment in innovation or a rise in innovation costs. The latter may arise when it becomes harder to invent new product lines. With the new innovation function the dynamics of product span, (1.13), become:

$$\dot{r}_i = \kappa\phi(\iota_i) - \theta r_i, \text{ for all } t \geq 0. \tag{1.25}$$

In this case the first-order condition of the optimal control problem (1.16) becomes:

$$\lambda_i \kappa \phi'(\iota_i) = 1, \tag{1.26}$$

while the differential equation (1.19) does not change. From (1.26) we obtain the investment level ι_i as an increasing function of $\kappa\lambda_i$, which we express as $\iota_i(\kappa\lambda_i)$. This is the same $\iota_i(\cdot)$

function that we had before. Substituting this function into (1.25) yields the autonomous differential equation:

$$\dot{r}_i = \kappa\phi\left[\iota_i\left(\kappa\lambda_i\right)\right] - \theta r_i.$$

The steady state of this differential equations is characterized by:

$$\kappa\phi\left[\iota_i\left(\kappa\lambda_i\right)\right] = \theta r_i,$$

while the second steady state equation, (1.22), does not change, because the differential equation (1.19) remains the same. For $\kappa = 1$, the steady state and the dynamics depicted in Figure 1.1 remain the same.

Now consider an increase in κ, representing a decline in the costs of inventing new product lines. Since $\kappa\phi\left[\iota_i\left(\kappa\lambda_i\right)\right]$ is increasing in κ, this leads to a downward shift of the $\dot{r}_i = 0$ curve without changing the $\dot{\lambda}_i = 0$ curve. As a result, λ_i declines on impact to a new saddle path, starting transition dynamics with declining values of λ_i and rising values of r_i. This process takes place in every large firm, leading to a new steady state in which every large firm has a larger product span, a larger market share and a higher markup. The average markups μ_{av} and μ^s_{av} rise during the transition and they are higher in the new steady state. The flow of aggregate utility also rises during this transition and is higher in the new steady state. The flow utility rises because profits net of investment costs rise while the price index P remains the same. We therefore have

Proposition 4. *Suppose that every large firm i is in steady state and $\bar{r} > 0$ at all times. Then a decline in the cost of innovation, i.e., an increase in κ, leads all large firms to expand product ranges, raise their market shares and raise their markups. Contemporaneously, the average markups μ_{av} and μ^s_{av} increase and so does the aggregate flow of utility.*

We next turn to changes in the marginal costs of production and the cost of entry of single-product firms. As is evident from (1.21) and (1.22), such changes impact the new steady state through the function $\Gamma_i\left(r_i\right)$ only. A change that raises $\Gamma_i\left(r_i\right)$ shifts upward the $\dot{\lambda}_i = 0$ curve in Figure 1.1. After the impact effect, which results from the upward jump in

λ_i, the dynamic process leads to a gradual widening of the span of products and increases in the markup and profits net of investment costs. In contrast, a change that reduces $\Gamma_i(r_i)$ shifts downward the $\dot{\lambda}_i = 0$ curve. After the downward jump of λ_i on impact, the dynamic process then leads to a gradual narrowing of the span of products and declines in markups and profits net of investment costs.

First, consider a decline in a_i, resulting from a technical improvement in the firm's technology. We show in the Appendix that the impact of a_i on Γ_i can be expressed as:

$$\begin{aligned}\hat{\Gamma}_i &= -(\sigma-1)\hat{a}_i + \left(\frac{\partial \Gamma_i}{\partial s_i}\frac{s_i}{\Gamma_i}\right)\left(\frac{\partial s_i}{\partial a_i}\frac{a_i}{s_i}\right)\hat{a}_i \\ &= \frac{(\sigma-1)s_i^2\delta^2 - (\sigma-\delta s_i-1)^2\left(\sigma^2-\delta^2 s_i^2\right)}{\left[(\sigma-\delta s_i-1)\sigma + s_i^2\delta^2\right]^2}(\sigma-1)\hat{a}_i.\end{aligned} \tag{1.27}$$

The relationship between a_i and Γ_i portrayed by this equation does not depend on the cost structure of other firms. Moreover, it implies that a decline in a_i shifts upward the $\dot{\lambda}_i = 0$ curve if and only if:

$$(\sigma-\delta s_i-1)^2\left(\sigma^2-\delta^2 s_i^2\right) > (\sigma-1)s_i^2\delta^2. \tag{1.28}$$

The potential ambiguity of the response of Γ_i to changes in a_i results from the existence of two channels through which the marginal cost impacts the profitability of a new variety (the marginal profitability of r_i), as can be seen from (1.20). A decline in a_i raises the marginal profitability of a new variety, Γ_i, for a given market share, s_i, due to cost savings in production. But, a decline in a_i makes firm i more competitive, thereby raising its market share, as shown in (1.11). A rise in the firm's market share reduces in turn the profitability of a new variety, as we show formally in the Appendix (see (A.1) in the Appendix). It follows that the shift of the $\dot{\lambda}_i = 0$ curve depends on the strength of these two effects: if the response of the market share dominates, the curve shifts down; and if the response of the market share does not dominate, the curve shifts up. The strength of the market share effect depends in turn on the firm's initial size. For low values of s_i the impact through the market share channel is small, and (1.28) is satisfied. But (1.28) is less likely to be satisfied the larger s_i is, because the left-hand side of this inequality is declining in s_i while the right-hand side

is increasing. This leads to the following

Lemma 1. *If* $(\sigma-\delta-1)^2\left(\sigma^2-\delta^2\right) > (\sigma-1)\delta^2$, *then (1.28) is satisfied for all market shares* $s_i \epsilon [0,1]$. *And if* $(\sigma-\delta-1)^2\left(\sigma^2-\delta^2\right) < (\sigma-1)\delta^2$, *then there exists a market share* $s^o \epsilon (0,1)$, *defined by:*

$$(\sigma-\delta s^o-1)^2\left[\sigma^2-\delta^2\left(s^o\right)^2\right] = (\sigma-1)\left(s^o\right)^2\delta^2,$$

such that (1.28) is satisfied for $s_i < s^o$ *and violated for* $s_i > s^o$.

Given the assumption $\sigma > \varepsilon > 1$, the inequality $(\sigma-\delta-1)^2\left(\sigma^2-\delta^2\right) > (\sigma-1)\delta^2$ is satisfied when ε is close to σ and violated when ε is close to one (recall that $\delta = \sigma - \varepsilon$). We therefore have

Proposition 5. *Suppose that firm i is in steady state and* $\bar{r} > 0$ *at all times. Then a decline in* a_i *triggers an adjustment process that gradually raises* r_i *as well as i's markup and profits net of investment costs if either* $(\sigma-\delta-1)^2\left(\sigma^2-\delta^2\right) > (\sigma-1)\delta^2$ *or* $(\sigma-\delta-1)^2\left(\sigma^2-\delta^2\right) < (\sigma-1)\delta^2$ *and* $s_i < s^o$, *where* s^o *is defined in Lemma 1. Otherwise, this technical improvement triggers an adjustment process that gradually reduces* r_i *while i's markup and profits net of investment costs decline gradually after increasing on impact.*

Using these results, we can examine the dynamics of firm i's market share. Since on impact the span of products does not change (r_i is a state variable), (1.11) implies that the decline in the marginal cost raises firm i's market share on impact. Moreover, if the adjustment process leads to a gradual expansion of its product span, i's market share rises over time until it reaches a new steady state. In this case the firm has a larger market share in the new steady state. If, however, the adjustment process leads to a narrowing of the firm's product span, then (1.11) implies that the initial upward jump in firm i's market share is followed by a gradual decline in its market share. A question then arises whether this firm's market share is larger or smaller in the new steady state. We prove the following

Proposition 6. *Suppose that firm i is in steady state and* $\bar{r} > 0$ *at all times. Then a decline*

in a_i triggers an adjustment process that raises s_i in the new steady state.

PROOF:

We have shown that the market share is larger in the new steady state when the adjustment process involves expansion of the firm's product span. It therefore remains to show that this is also true when the adjustment process involves contraction of the product span. To this end note that a decline in r_i on the transition path is triggered by a decline in the marginal profitability of r_i in response to a decline in a_i, which leads in turn to a downward shift in the $\dot{\lambda}_i = 0$ curve in Figure 1.1. In this case the new steady state has a lower r_i as well as a lower λ_i. Next note from the steady state condition (1.22) that a lower λ_i implies a lower Γ_i. Recall, however, that for a constant s_i a fall in a_i raises Γ_i, and therefore Γ_i can be lower in the new steady state only if s_i is higher. In sum, independently of whether a decline in a_i shifts upward or downward the $\dot{\lambda}_i = 0$ curve, the market share s_i is larger in the new steady state.

This result yields the following

Corollary 1. *Consider an economy in steady state with active single-product firms. Then large firms with lower marginal costs have larger market shares.*

The dynamic patterns of the market share that have been unveiled by this analysis are depicted in Figure 1.2, where s_i^1 is the market share in the initial steady state. First, the market share jumps up to s_i^{im} on impact when a_i declines. Afterward, the market share rises continuously until it reaches s_i^2, as portrayed by the upper curve, or it declines continuously until it reaches s_i^3, as portrayed by the lower curve. In both cases the new steady state market share exceeds s_i^1. The former case applies when $(\sigma - \delta - 1)^2 \left(\sigma^2 - \delta^2\right) > (\sigma - 1)\,\delta^2$ or $(\sigma - \delta - 1)^2 \left(\sigma^2 - \delta^2\right) < (\sigma - 1)\,\delta^2$ and $s_i^1 < s^o$, and the latter case applies otherwise.

These results suggest three possible steady state patterns for the relationship between a_i and r_i in the cross section of multi-product firms: lower-cost firms have larger product spans, lower-cost firms have smaller product spans, or the relationship between marginal costs and product spans has an inverted U shape. The first pattern holds for all marginal

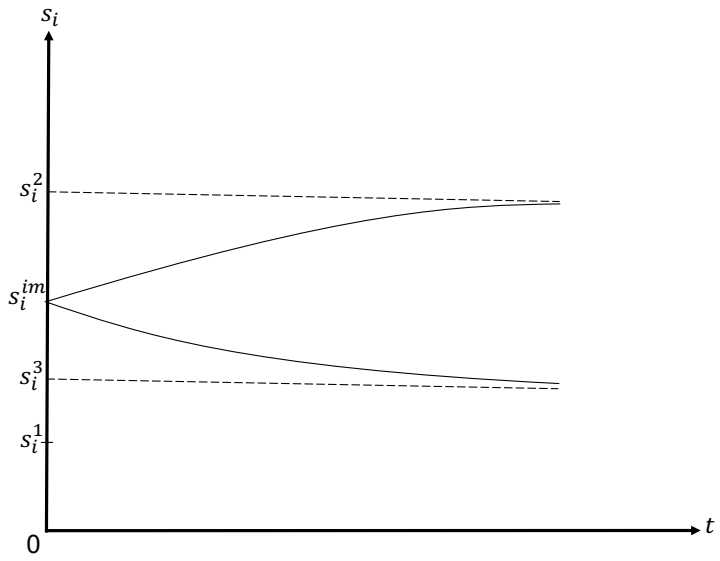


Figure 1.2: *Dynamics of the market share in response to a decline in the marginal cost a_i*

cost structures when $(\sigma-\delta-1)^2\left(\sigma^2-\delta^2\right) > (\sigma-1)\delta^2$. In the opposite case, when $(\sigma-\delta-1)^2\left(\sigma^2-\delta^2\right) < (\sigma-1)\delta^2$, there exist high values of a_i at which $s_i < s^o$, and among firms with marginal costs in this range those with lower marginal costs have larger product spans. Moreover, there exist low values of a_i at which $s_i > s^o$, and among firms with such low marginal costs lower-cost firms have smaller product spans. Combining these results we have

Proposition 7. *Consider an economy in steady state with active single-product firms. Then, in the cross section of multi-product firms r_i is declining in s_i, rising in s_i, or rising in s_i among firms with low market shares and declining in s_i among firms with high market shares.*

Combining this Proposition with Corollary 2, we note that our model raises the possibility of an inverted-U relationship between labor productivity, as measured by $1/a_i$, and the number of product lines, r_i. We now show that this prediction is not only a theoretical possibility, but that there is suggestive evidence for such a relationship in the Compustat data set. To this end we collected data on revenue, employment, the number of sectors in which a firm operated and the number of segments in which a firm operated, all for

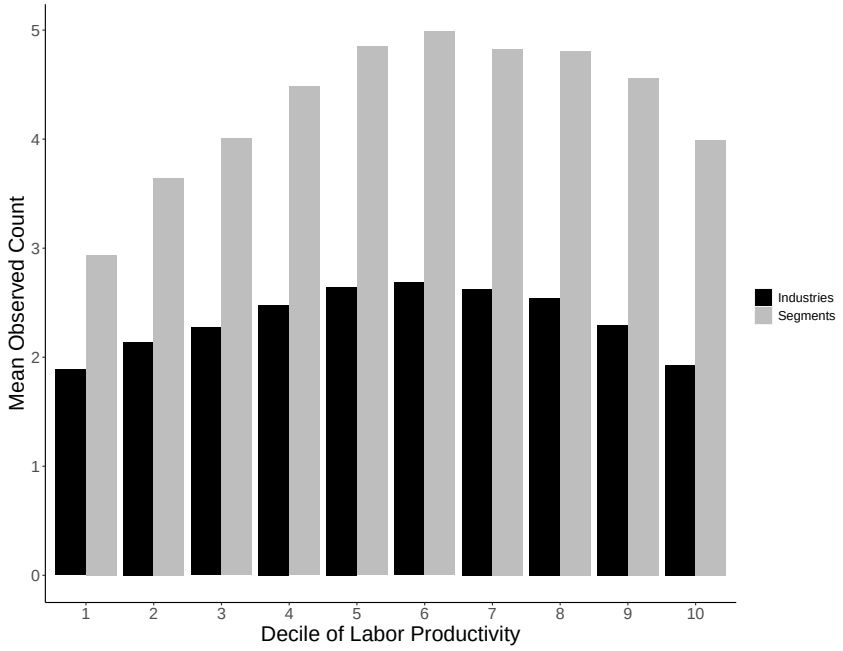


Figure 1.3: *Average Number of Product Lines vs. Labor Productivity Deciles*

2018.[15] We computed labor productivity as revenue per worker and we treat the number of segments as a proxy for the number of product lines. As a robustness check, we also consider the number of industries in which a firm operated as a proxy for the number of its product lines.[16] Figure 1.3 depicts the relationships between our two proxies for r_i and our proxy for $1/a_i$. On the horizontal axis the firms are divided into deciles, based on their labor productivity. On the vertical axis we report the mean number of segments and the mean number of industries in each decile. As is evident, these relationships exhibit an inverted-U.

To further examine these relationships, we regressed the number of segments or the

[15]Compustat data is produced by Standard and Poor's (S&P). The data was accessed through Wharton Research Data Services. This service and the data available thereon constitute valuable intellectual property and trade secrets of WRDS and/or its third-party suppliers.

[16]About 70% of the firms in the Compustat database breakdown the company into segments through Compustat Segments Data. Firms are able to distinguish between business segments, geographic segments, operating segments, state segments. This data is self-reported and thus is not standardized, but is still widely used. We focus on the number of business segments a company lists as a proxy for the number of product lines. Within each business segment the firm can list up to two SIC codes in which the business segment operates. The total number of unique SIC codes listed across business segments is what we define as the number of industries in which a firm operates. This is our second proxy for the number of product lines.

Figure 1.4: *Number of Segments vs. Labor Productivity*

number of sectors in which a firm operates on a second-order polynomial of the log of labor productivity. We report in the Appendix the resulting OLS estimates. The coefficient on log labor productivity is positive and the coefficient on the log of labor productivity squared is negative in both case. Moreover, all four coefficients are significantly different from zero. Figure 1.4 plots the data points that we have used (more than 4,000 observations) as well as the fitted quadratic curve. The first thing to note is that there are many firms with similar numbers of segments and different labor productivity levels, especially when the number of segments is low. Nevertheless, the estimated curve has the shape of an inverted-U. We report in the Appendix a similar graph for the number of industries in which a company operates. In conclusion, while we view this paper as a theoretical contribution, we have also provided suggestive evidence for the inverted-U curve predicted by our model.

1.5.1 Costs of Single-Product Firms

We next examine the impact of the cost structure of single-product firms. As is evident from (1.9), a decline in either the marginal cost or the entry cost of single-product firms reduces

the price index P, thereby raising the competitive pressure in the economy. How do the large firms respond to this rise in competition? To answer the question, suppose that all firms are in steady state. Equation (1.20) implies:

$$\hat{\Gamma}_i = \delta \hat{P} + \left(\frac{\partial \Gamma_i}{\partial s_i} \frac{s_i}{\Gamma_i} \right) \left(\frac{\partial s_i}{\partial P} \frac{P}{s_i} \right) \hat{P}. \tag{1.29}$$

A decline in the price index P elevates the competitive pressure on every large firm and reduces the marginal value of its product span, r_i. Accordingly, the first term on the right-hand side of this equation is negative when $\hat{P} < 0$. In response, firm i reduces its price and market share (see (1.10) and (1.11)) and the fall in market share raises the marginal value of r_i. For this reason the second term on the right-hand side is positive when the price index declines. It follows that a decline in P shifts the $\dot{\lambda}_i = 0$ curve downward in Figure 1.1 if the competition effect dominates and upward if the market share effect dominates. Using (1.11), it is evident that for $\varepsilon \to 1$ (1.29) is similar to (1.27), except for the opposite sign on their right-hand sides. Therefore, in this case a decline in P shifts down the $\dot{\lambda}_i = 0$ curve if and only if a decline in a_i shifts it up. Under these conditions a lower P may lead to a lower or higher value of r_i in steady state, and moreover, its impact may vary across firms with different marginal costs and therefore different market shares s_i. For $\varepsilon \to 1$ the inequality $(\sigma - \delta - 1)^2 \left(\sigma^2 - \delta^2\right) > (\sigma - 1)\, \delta^2$ is violated, implying that there exists an s_P^o, such that the decline in P shifts the $\dot{\lambda}_i = 0$ curve down for $s_i < s_P^o$ and up for $s_i > s_P^o$. In this case a rise in the competitive pressure shrinks the product span of multi-product firms with $s_i < s_P^o$ and expands the product span of multi-product firms with $s_i > s_P^o$. As a result, the gaps in market shares between large and small multi-product firms widens, thereby increasing the inequality in the size distribution of firms.[17] Alternatively, for $\varepsilon \to \sigma > 1$ the inequality $(\sigma - \delta - 1)^2 \left(\sigma^2 - \delta^2\right) > (\sigma - 1)\, \delta^2$ is always satisfied, implying that for ε close to σ the competition effect dominates the market share effect. Consequently, the $\dot{\lambda}_i = 0$ curve shifts down for all multi-product firms, decreasing their product span, when the price index decreases.

[17] From (1.11), $\hat{s}_i - \hat{s}_j = \left(\hat{r}_i - \hat{r}_j\right) / \left[1 + (\sigma - 1)\, \beta_i\right]$. Therefore $\hat{s}_i > \hat{s}_j$ if and only if $\hat{r}_i > \hat{r}_j$.

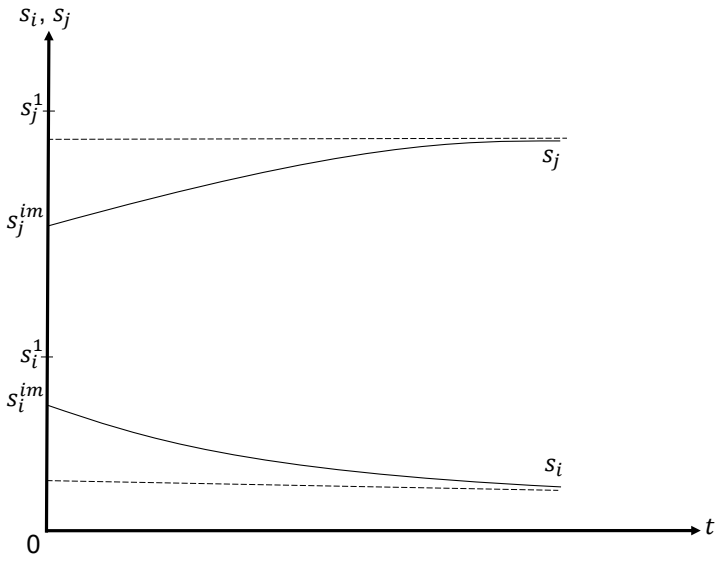


Figure 1.5: *Dynamics of market shares in response to a decline in P*

Finally, note that a decline in P reduces the steady state market share of every large firm. This is clearly the case when every firm's product span declines, because in this case declines in both P and r_i diminish the market share (see (1.11)). Alternatively, for a firm that expands its steady state r_i, the value of λ_i is higher in the new steady state (see (1.21)). Therefore, this firm's Γ_i is also larger in the new steady state (see (1.22)). But the direct impact of the decline in P on Γ_i is negative, and therefore s_i has to be smaller for Γ_i to be larger. We, therefore, have

Proposition 8. *Consider an economy in steady state with $\bar{r} > 0$ at all times. Then, a technical improvement that reduces either f or $\bar{a}$ may raise r_i in the new steady state for all i, reduce r_i for all i, or reduce r_i of the small multi-product firms and raise r_i of the large multi-product firms. Nevertheless, s_i is smaller in the new steady state for all large firms i.*

Figure 1.5 depicts the dynamics of two firms, i and j, for the case in which $s_i < s_P^o$ and $s_j > s_P^o$, where s_P^o is the cutoff market share for the opposite firm dynamics. Firm i starts with $s_i = s_i^1$ while firm j starts with $s_j = s_j^1$. In both firms the market share jumps down on impact as a result of the decline in P, to s_i^{im} and s_j^{im}, respectively. After that, the market share of the smaller firm declines while the market share of the larger firm rises. Yet in both

cases, the market share is lower in the new steady state.

Gutierrez and Philippon (2019) find that the elasticity of the number of firms with respect to Tobin's Q declined during 1995-2010. They argue that this resulted from increased entry costs due to regulation rather than due to technological developments or financial frictions. In our model an increase in f generates the above described dynamics independently of the source of variation in the fixed cost of entry. According to Proposition 8, an increase in f raises the long-run market share of all large multi-product firms and reduces the joint market share of the small single product firms. Yet, it may have an uneven impact on the span of products of the large firms. That is, it may increase the number of product lines of the smaller multi-product firms and reduce the number of product lines of the large ones, thereby flattening the relationship between labor productivity (i.e., $1/a_i$) and product span.

1.6 Optimal Allocation

We study in this section the optimal allocation and discuss policies that support it in a decentralized equilibrium. Recall that the interest rate is constant and equal to ρ. Therefore the optimal allocation is obtained by maximizing the present value of the utility flows $\int_0^\infty e^{-\rho t} u(t) dt$, where u is given by (1.1). We characterize the optimal allocation in two stages. First, we solve a static optimal allocation for every point in time. Then, in stage two, we solve a dynamic optimal allocation problem that uses the static solution at each point in time. This two-stage procedure provides useful intuition and makes it easier to characterize optimal policies in a market economy.

At a point in time t consumption of the homogeneous good is equal to

$$x_0(t) = l - \left\{ \bar{r}(t) \left[f + \bar{a}\bar{x}(t) \right] + \sum_{i=1}^{m} \left[\iota_i(t) + r_i(t) a_i x_i(t) \right] \right\}.$$

The term in the curly brackets represents labor used in the production of varieties of the differentiated product by single- and multi-product firms, plus the entry resource cost of single-product firms, plus investment in multi-product firms for the expansion of their product spans, where all variables have the same meaning as in the previous section. Using

the definition of the real consumption index X in (1.2),

$$X(t) = \left[\bar{r}(t) \bar{x}(t)^{\frac{\sigma-1}{\sigma}} + \sum_{i=1}^{m} r_i(t) x_i(t)^{\frac{\sigma-1}{\sigma}} \right]^{\frac{\sigma}{\sigma-1}}. \tag{1.30}$$

Substituting these equations into (1.1) we obtain the flow of utility at time t

$$u(t) = l + \frac{\varepsilon}{\varepsilon-1} X(t)^{\frac{\varepsilon-1}{\varepsilon}} - \left[\bar{r}(t) \bar{a} \bar{x}(t) + \sum_{i=1}^{m} r_i(t) a_i x_i(t) \right] - \bar{r}(t) f - \sum_{i=1}^{m} \iota_i(t).$$

The term in the square brackets on the right-hand side of this equation represents variable labor costs of producing the real consumption $X(t)$. To achieve optimality these labor costs have to be minimized subject to (1.30), yielding the cost function $C\left[\bar{r}(t), \{r_i(t)\}_{i=1}^{m}\right] X(t)$, where

$$C\left[\bar{r}(t), \{r_i(t)\}_{i=1}^{m}\right] = \left[\bar{r}(t) \bar{a}^{1-\sigma} + \sum_{j=1}^{m} r_j(t) a_j^{1-\sigma} \right]^{\frac{1}{1-\sigma}} \tag{1.31}$$

is the unit labor cost of $X(t)$. Substituting this result into the utility flow yields:

$$u(t) = l + \frac{\varepsilon}{\varepsilon-1} X(t)^{\frac{\varepsilon-1}{\varepsilon}} - C\left[\bar{r}(t), \{r_i(t)\}_{i=1}^{m}\right] X(t) - \bar{r}(t) f - \sum_{i=1}^{m} \iota_i(t).$$

Now note that optimality requires to choose $X(t)$ so as to maximize this utility flow, which yields $X(t) = C\left[\bar{r}(t), \{r_i(t)\}_{i=1}^{m}\right]^{-\varepsilon}$ and the flow of utility

$$u(t) = l + \frac{1}{\varepsilon-1} C\left[\bar{r}(t), \{r_i(t)\}_{i=1}^{m}\right]^{1-\varepsilon} - \bar{r}(t) f - \sum_{i=1}^{m} \iota_i(t). \tag{1.32}$$

Next observe that $\bar{r}(t)$ has to be chosen so as to maximize (1.32) subject to (1.31), yielding the first-order condition

$$\frac{1}{1-\sigma} C\left[\bar{r}(t), \{r_i(t)\}_{i=1}^{m}\right]^{\delta} \bar{a}^{1-\sigma} = f, \ \delta = \sigma - \varepsilon > 0.$$

We expressed this as an equality condition, assuming that in every period the left-hand side of this equation is larger than the right-hand side for $\bar{r}(t) = 0$; otherwise, it is not desirable to have active single-product firms. For example, when the entry cost of single-product firms, f, is very high, it maybe optimal to forgo their services. For comparison with our market results, where we assumed that the small firms are viable, we assume that the

fixed cost f is small enough so that $\bar{r}(t) > 0$ for all t on the optimal path. In this case the first-order condition for the choice of $\bar{r}(t)$ is satisfied with equality at every point in time, implying that the unit cost $C(\cdot)$ is constant on the dynamic path and equal to C^*, implicitly defined by

$$\frac{1}{1-\sigma} C^{*\delta} \bar{a}^{1-\sigma} = f. \tag{1.33}$$

In what follows we use asterisks to denote optimal values of endogenous variables. This result implies that the real consumption index X is also constant on the dynamic path and $X(t) = X^* = C^{*-\varepsilon}$ for all t. Finally, from the minimization problem that was used to derive the cost function $C(\cdot)$ we obtain the optimal output quantities:

$$\bar{x}^* = C^{*\delta} \bar{a}^{-\sigma} = (\sigma - 1)\bar{a}^{-1} f, \tag{1.34}$$

$$x_i^* = C^{*\delta} a_i^{-\sigma} = (\sigma - 1)\bar{a}^{\sigma-1} a_i^{-\sigma} f,\ i = 1, 2, ..., m. \tag{1.35}$$

Evidently, the optimal output level of every variety is constant and a firm's output level is larger the lower its marginal cost. These findings are summarized in

Proposition 9. *Consider the optimal allocation in an economy that has a low entry cost f that secures $\bar{r}^* > 0$ for all t. Then: (i) the output levels $\bar{x}^*$ and $\{x_i^*\}_{i=1}^m$ are constant on the optimal path; (ii) the unit cost C^* and the real consumption index X^* are constant on the optimal path and $X^* = C^{*-\varepsilon}$.*

Using these results, we can express the flow of utility as a function of product spans of the multi-product firms and their investment levels. From (1.31) and (1.32) and the result that $C(\cdot) = C^*$ on the dynamic path, the flow of utility satisfies:

$$u\left[\{r_i(t)\}_{i=1}^m, \{\iota_i(t)\}_{i=1}^m\right] = l + \frac{1}{\varepsilon - 1} C^{*1-\varepsilon} - \left[C^{*1-\sigma} - \sum_{i=1}^m r_i(t)\, a_i^{1-\sigma}\right] \bar{a}^{\sigma-1} f - \sum_{i=1}^m \iota_i(t). \tag{1.36}$$

The term in the square brackets shows that larger product spans of multi-product firms call for fewer single-product firms in order to ensure a constant value $C(\cdot) = C^*$. This lowers entry costs of the small firms, thereby saving resources and raising welfare. There

is, however, a tradeoff: the growth of product span of large firms requires investment in innovation, which reduces consumption and welfare. This tradeoff is optimized in the dynamic problem that we solve next.

The dynamic optimal allocation problem can be formulated as follows:

$$\max_{\left[\{r_i(t)\}_{i=1}^m,\{\iota_i(t)\}_{i=1}^m\right]_{t\geq 0}} \int_0^\infty e^{-\rho t} u\left[\{r_i(t)\}_{i=1}^m, \{\iota_i(t)\}_{i=1}^m\right] dt$$

subject to (1.13), (1.36), $r_i(0) = r_i^0$, and transversality conditions to be described below. In this problem the state variables are $\{r_i(t)\}_{i=1}^m$ while the control variables are $\{\iota_i(t)\}_{i=1}^m$. Dropping t in the notation of time-dependent variables, the current-value Hamiltonian is:

$$\mathcal{H} = l + \frac{1}{\varepsilon - 1} C^{*1-\varepsilon} - \left(C^{*1-\sigma} - \sum_{i=1}^m r_i a_i^{1-\sigma} \right) \overline{a}^{\sigma-1} f - \sum_{i=1}^m \iota_i + \sum_{i=1}^m \lambda_i \left[\phi(\iota_i) - \theta r_i\right],$$

where λ_i is the co-state variable of constraint (1.13). The first-order conditions of this problem are:

$$\frac{\partial \mathcal{H}}{\partial \iota_i} = -1 + \lambda_i \phi'(\iota_i) = 0, \tag{1.37}$$

$$-\frac{\partial \mathcal{H}}{\partial r_i} = -\left(\frac{\overline{a}}{a_i}\right)^{\sigma-1} f + \theta \lambda_i = \dot{\lambda}_i - \rho \lambda_i, \tag{1.38}$$

and the transversality conditions are:

$$\lim_{t\to\infty} e^{-\rho t} \lambda_i(t) r_i(t) = 0.$$

In addition, the optimal path of $\{\iota_i, r_i\}_{i=1}^m$ has to satisfy the differential equations (1.13). Equation (1.38) yields the autonomous differential equation

$$\dot{\lambda}_i = (\rho + \theta)\lambda_i - \left(\frac{\overline{a}}{a_i}\right)^{\sigma-1} f.$$

A second differential equation is obtained from (1.13) and (1.37):

$$\dot{r}_i = \phi\left[\iota_i(\lambda_i)\right] - \theta r_i.$$

These equations generate transition dynamics similar to Figure 1.1, except that now the

$\dot{\lambda}_i = 0$ curve is horizontal and therefore λ_i is constant during the transition and equal to:

$$\lambda_i^* = \frac{1}{\rho + \theta} \left(\frac{\overline{a}}{a_i} \right)^{\sigma - 1} f. \tag{1.39}$$

This implies that the dynamic system travels on the $\dot{\lambda}_i = 0$ curve and it satisfies the transversality conditions. Since all multi-product firms share the same $\dot{r}_i = 0$ curve but their $\dot{\lambda}_i = 0$ curves differ according to a_i, it follows that firms with lower marginal costs a_i end up with larger product spans in the steady state. In other words, in the steady state of the optimal allocation there is a *monotonically* decreasing relationship between marginal cost and product span in the cross section of multi-product firms. Evidently, an inverted-U curve relationship between these variables arises only in a distorted economy. Finally, note that due to the fact that the co-state variable λ_i^* is constant on the optimal path, so is investment in innovation ι_i^*, as can be seen from the first-order condition (1.37). These findings are summarized in the following

Proposition 10. *Consider the optimal allocation in an economy that has a low entry cost f that secures $\bar{r}^* > 0$ for all t. Then: (i) investment in innovation ι_i^* is constant on the optimal path and larger the smaller is a multi-product firm's marginal cost; (ii) in steady state multi-product firms with lower marginal costs have larger product spans; (iii) if $r_i(0) = r_i^0$ is smaller than the optimal steady state value of r_i^* for all i, then the product span of every large multi-product firm rises and the number of small single-product firms declines on the optimal path.*

To decentralize the optimal allocation, it is necessary to subsidize consumer purchases of every firm's varieties so as to ensure consumer prices that equal marginal costs of production. In addition, every large firm's operating profits have to be taxed to ensure that the firm perceives a constant marginal value of product spans on its entire dynamic path, yet no policy is required to modify entry incentives of single-product firms. We provide in the Appendix a full characterization of these policies. One important feature of the optimal policies is that the subsidy to consumers on purchases of products supplied by single-product firms does not vary over time, while subsidies to consumer purchases of

products supplied by large multi-product firms have to vary over time on the transition path. Second, operating profits of large firms have to by taxed in order to induce them to engage in optimal investment in innovation. When the initial product span of a firm is below the optimal steady state value, this firm expands its product span over time. In this case the time pattern of the optimal corporate tax depends on the relative size of the elasticity of substitution σ and the sectoral elasticity of demand ε. If $\sigma > 2\varepsilon$ there exists a market share $s_c = \sigma/2\,(\sigma - \varepsilon)$ such that the tax rate is rising if the share of consumer spending on the firm's products is lower than s_c and the tax rate is declining if the share of consumer spending on the firm's products is larger than s_c. In the opposite case, $\sigma < 2\varepsilon$, the corporate tax rate always rises for a firm that grows its product span. Derivation of these results are provided in the Appendix.

1.7 Conclusion

We have developed a parsimonious model of industry evolution, in which large multi-product firms grow via investment in new product lines. While these firms are oligopolies, they face competitive pressure from small single-product firms that engage in monopolistic competition. Our model generates time patterns of markups, concentration, and labor shares that are consistent with the data. Moreover, it predicts rich patterns for the cross-section of firms. In particular, it predicts an inverted-U relation between labor productivity and product span, for which we provide supportive evidence. It also predicts that rising competitive pressure from small single-product firms flattens the cross-sectional relationship between labor productivity and product span among the large multi-product firms.

We also characterized the optimal allocation and compared it to the market outcome. In the optimal allocation there is a monotonically increasing relationship between labor productivity and product span, which implies that the inverted-U relationship is caused by misallocation.

Although this study consists of a theoretical contribution, we believe that our model delivers valuable insights into industry dynamics that can be empirically studied. There are

few data sets containing information on product span of individual firms, and these data are mostly confidential. Nevertheless, we hope that the predictions of our model will eventually be examined with some of the existing rich data sets. Finally, we show in the appendix how to construct an aggregate economy with a continuum of industries of the type studied in this paper. This model economy can be used to study various macroeconomic issues, including economic growth.

Chapter 2

What You See is What You Get: Local Labor Markets and Skill Acquisition

2.1 Introduction

Alfred Marshall's statement that "when an industry has thus chosen a locality for itself...the mysteries of the trade are no mysteries; but are as it were in the air..." (Marshall (1890)) has become the jumping off point for modern thinking on agglomeration[1]. Recently, it has formed the basis for a barrage of papers that have been written on the potential for place based policies in response to recent trends of regional divergence and the concentration of skilled labor documented in Berry and Glaeser (2005) and Moretti (2013). Policy prescriptions based on these research efforts vary widely from relatively laissez faire (Glaeser and Gottlieb (2008)) to more interventionist (Rossi-Hansberg *et al.* (2019)). The common question being whether the benefits of agglomeration stemming from attracting skilled workers to a particular area outweigh the costs including the agglomeration losses in the region of origin (i.e. brain drain). These papers typically incorporate a static spatial model focusing on the productive benefits of concentrating skilled workers.

[1]See for instance Ellison *et al.* (2010), Davis and Dingel (2019), Kantor and Whalley (2014), Kline and Moretti (2014), Zacchia (2020)

This paper builds off of this scaffolding in order to look more closely at the impact of local labor markets on skill acquisition. In this sense, I am addressing the rest of Alfred Marshall's influential proclamation: "so great are the advantages which **people following the same skilled trade** get from near neighbourhood to one another... children learn many of them unconsciously". The implication here is that people seem to be naturally predisposed to follow in the footsteps of the trades that people around them are engaged in. There is a positive externality exerted when people are surrounded by in-demand skills which leads them to pursue an education with high expected returns. The dynamic impact of concentrating labor in skill hubs is that it increases the likelihood that people will acquire in demand skills in those select few labor markets at the expense of reducing the likelihood elsewhere. If the impact of local labor markets on skill acquisition is concave, then this will reduce the overall pool of in-demand skills in the economy.

In order to encourage future research along this vein, this paper provides evidence of the importance of understanding the impact of local labor markets on skill choice and provides the first steps in analyzing its importance relative to the standard focus on agglomeration. I first highlight the disparate outcomes associated with different majors, which goes a long way towards explaining findings that students have a difficult time predicting expected incomes by major as documented in Wiswall and Zafar (2015). In response to this uncertainty, students tend to be biased towards the majors and occupations that they are able to observe in their local labor markets. The goal here is to stress the fact that with the variation across majors being nearly as large as the overall college skill premium, bias towards particular majors/degrees can have a major impact on future earnings. Spatial divergence by major and occupation is also exacerbated by recent trends towards more limited migration.

However, even if there is bias towards local labor market skills beyond the local wage, this may not be sufficient to call for a more even spatial distribution of skills. First, it must be the case that the impact is concave such that the benefits of concentrating skills are smaller than the benefits of more evenly distributing skills. Secondly, concentrating skills may still create a high enough wage premium to overcome local skill bias. In order to consider this

tradeoff more carefully, I develop a structural spatial model with two key externalities. The first is the standard agglomeration externality by which an increasing fraction of skilled labor in a local labor market increases the productivity of local workers. Second, I introduce a new externality that comes from signaling the expected wage associated with particular skills across potential locations. Although I frame it in terms of wage signaling, the reduced form evidence I provide is isomorphic to a broad range of potential biases with the common feature that students tend to be more likely to get the skills demanded in their local labor market beyond features related to the wage.

Rather than calibrate the full structural model, I make several clarifying assumptions to allow me to present some suggestive reduced form evidence of the importance of this channel. This analysis is not meant to end the discussion through precise estimates of biases due to local labor market conditions but merely to introduce it. In my simplified model using decennial Census data for the years 1980,1990 and 2000, I find evidence that both the agglomeration and signaling externalities are of similar magnitudes with the important feature that concentrating skills has a heterogeneous impact across locations through the signaling externality. The effect of a more even distribution of skills across locations is convex in the initial skill concentration across locations. MSAs with low skill levels before redistribution experience large positive gains relative to initially high skill MSAs which experience small losses. Overall, this suggests that policies which affect the distribution of skills across space should consider not just the impact on productivity but on how the local labor market will impact skill acquisition more broadly.

Related Literature: This paper speaks to three distinct strands in the literature. Most directly this relates to the literature on place based policies. As alluded to above, many papers have taken the presence of agglomeration as inspiration to assess whether the government should be involved in shifting where people live. Traditionally the justification for concentrating workers in a particular location has hinged on whether the benefits of agglomeration are convex (Glaeser and Gottlieb (2008)). More recently, the focus has been on whether skilled workers provide greater value to other skilled workers than to unskilled

workers. When defining skilled workers as those performing cognitive, non-routine (CNR) tasks, Rossi-Hansberg *et al.* (2019) finds that the optimal policy involves creating cognitive hubs by subsidizing CNR workers to move to large skilled cities and taxing non-CNR workers in those cities.

Similarly, Fajgelbaum and Gaubert (2020) find that the optimal policy involves increasing the concentration of skilled workers however, they achieve this by having both college educated and non-college educated workers moving away from the largest city. In this way all cities end up with a higher share of skilled workers. My baselines findings are roughly in line with this result. Note that this is the opposite of recent trends in the United States as described in Moretti (2012). This paper intends to add additional motivation to this debate by considering the downstream ramifications of the recent concentration of skilled workers in cognitive hubs on the future supply of skilled workers through signaling externalities.

This paper also connects to the vast literature on educational choice. The ground breaking paper by Wiswall and Zafar (2015) showed that students at NYU had inaccurate beliefs about expected wages for different majors. A major finding was that students adjust correctly when provided updated information, but that the majority of the variation in major choice was not explained by perceived wage or ability. Alternative factors which affect choice of major include exposure to role models (Porter and Serra (2020)) and industry experience (Boudreau and Marx (2019)). This is similar to the Bell *et al.* (2019) findings for innovation in which early exposure to inventors increased the likelihood of inventing in the corresponding field. One way to view my work is as an attempt to begin filling in the gap of the “taste residual" of educational choice.

Lastly, I hope to bring some additional perspective on the divergence in wages between skilled and unskilled workers. Autor (2019) summarizes a broad swath of recent research concerned with the future of work. It highlights the on-going trend towards higher skill premiums and higher skill share cities. The importance of acquiring in demand skills has grown dramatically as the middle class jobs that provide a substantial buffer between high and low income groups has been automated (Autor *et al.* (2006)). This growing inequality is

exacerbated by the relatively small gains in educational attainment made in recent years as discussed in Autor *et al.* (2020b). This problem is further exacerbated by the large discrepancies in wages by major Long *et al.* (2015) and differences in responsiveness to changes in the wage. My findings suggest that local bias will tend to push towards regional divergence and a disconnect between the aggregate skill premium and skill supply.

2.2 Suggestive Evidence

The importance of location **and** college major loom large in the economy. In this section, I hope to highlight the importance of location **on** major choice and how this has the potential to significantly distort labor market outcomes. These facts will be used as a starting point for the model that will be developed in Section 2.3.

2.2.1 Complexity in College Major Choice for Potential Income

It is especially important to study the determinants of skill investment because it is typically a one shot game with significant impact on lifetime income. This means that the plethora of potential mistakes won't be ironed out over time through repeated efforts. The decision of whether or not to go to college has because an increasingly significant decision as the skill premium has risen as documented in Katz and Murphy (1992). The decision is further complicated by the choice of major. The ratio of returns to high income vs. low income majors is on the same order as the college wage premium (Altonji *et al.* (2012)). Thus, the question of choosing the "right" major can be just as important as the decision of whether or not to go to college, and all of these decisions are typically made once in a lifetime.

This paper is not intended to do a deepdive on what drives the differences in returns (for more on this see Altonji *et al.* (2012) or Andrews *et al.* (2022)), but rather I focus on the spatial component of skill acquisition. However, it is useful to highlight the extreme differences in outcomes. Using data collected by IPUMS from the American Community Survey (ACS) for the years 2010-2018, I show in figure 2.1 the high level of variation of mean wage for workers from 25-55 (prime aged) with a college degree **across majors**. The highest

earning major is engineering with an average income over this time period of roughly $90K whereas the lowest earning major was library science with a mean total income of $36K. There is also considerable heterogeneity in terms of the risk associated with the different majors. The standard deviation ranging from $26K for library science to $82K for social sciences. In figure 2.1 we can see the trade off in terms of the mean log wage and the standard deviation of the log wage. The majors range drastically both in terms of mean and standard deviation as well as in the tradeoff between these dimensions. This evidence is roughly in line with Christiansen *et al.* (2007) which further argues that there is an efficient frontier in terms of the risk reward tradeoff for particular degrees. They take the fact that people choose degrees which are inside of the frontier as evidence for the importance of taste in educational attainment.

For the purposes of this paper, what is clear is that making an efficient educational choice requires substantial information about future earnings especially when there are concerns about the risk of particular majors. Wiswall and Zafar (2015) finds that risk aversion is quantitatively important. This is also true at the occupational level as in Dillon (2018) where workers are willing to accept lower earnings in order to avoid wage and employment risks. Especially given the fact that the expected outcomes for the average person in the population may be less important than signals about one's own potential outcome in evaluating a potential major, it is unsurprising that students have limited accuracy in terms of expected outcomes for the aggregate population.

2.2.2 College Major Choice is Consistent with Local Degree Distribution

Given the large distribution of potential outcomes based on college major, it is important to understand the underlying reasoning behind major choice. For instance, Wiswall and Zafar (2015) find that students respond to pecuniary factors but that the majority of the choice seems to be explained by taste parameters. In this paper, I argue that a large component of this apparent taste may be driven by growing up in a labor market with concentrations in particular skills.

Figure 2.1: *ACS Total Income by Major for Prime Aged College Graduates*

Note: This figure shows the mean and standard deviation of log of total annual wage and salary for employed workers with a college degree between the ages of 25 and 55. The data comes from the ACS for the years 2010-2018 and were acquired through IPUMS.

I provide reduced form evidence on the relationship between skill acquisition and local skill concentration using data from the Higher Education Research Institute (HERI) for the years 2000-2008. Specifically, I focus on the primary major stated by respondents of the True Freshman Survey, which has the broadest coverage in the survey data. This survey has roughly 300000 respondents each year and has been administered by over 1900 institutions. The data I look at includes both the stated major as well as the zipcode of each student's home address. This allows me to map respondents to their home counties and link the data to the ACS. Unfortunately, the ACS only includes degree field starting in 2009, so that in order to back out the number of prime aged workers by degree for the year 2000 I calculate the fraction of individuals in a particular occupation by degree for the years 2009-2018 and use this relationship to develop a proxy for local labor market skills. This allows me to approximate the number of people with a particular degree in a county using Census data

for the year 2000.

I show that there is a clear and statistically significant positive relationship between the fraction of workers with a particular degree and the fraction of freshman who declare that as their major at the home county level. This relationship is plotted in Figure 2.2. Clearly, degree attainment is positively correlated with local labor conditions. This relationship is particularly strong in business, health, social science and education while being much flatter in STEM degrees. This is what one would expect given the high level of public attention on the returns to STEM degrees and the importance of STEM education. However, other degrees such as business, social science and medicine which are on the higher end of the major earnings distribution are less vocally promoted. This implies that exposure to occupations that requires those degrees are more likely to affect the decision to major in those fields.

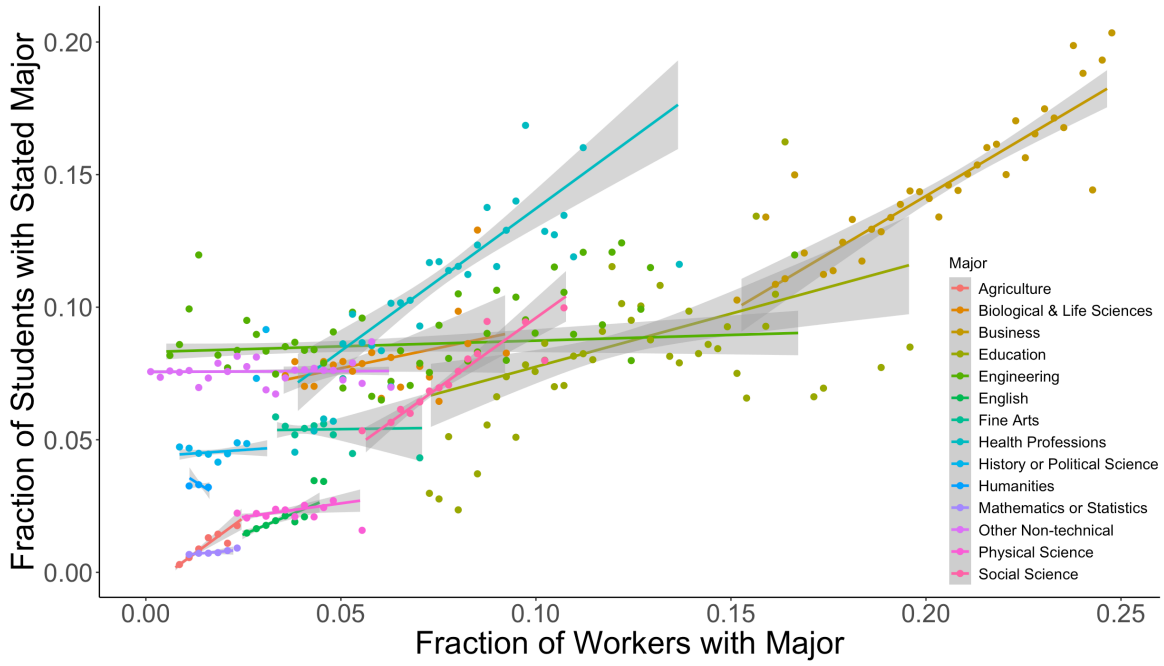


Figure 2.2: *Comparing Working Population Major Composition vs. College Student Stated Major*

Note: This figure plots the relationship between stated major in the TFS administered by the HERI and the fraction of workers in the students home county. The fraction of workers is approximated based on an occupational mapping from major to occupation from the ACS for years 2009-2018. The home county is based on a crosswalk from zipcode to county. In cases where there is a many-to-one match the student is mapped to each county.

Table 2.1: *Effect of Local Labor Market on Major Choice*

	Fraction Choosing that Major		
Fraction of Workers with Major in County	0.45^{***}	0.16^{***}	0.15^{**}
	(0.02)	(0.02)	(0.02)
Major FE	NO	YES	YES
County FE	NO	NO	YES
Obs	9395	9395	9395
R^2	49.6%	66.1%	67.4%

Robust standard errors clustered at the primary industry in parentheses.

$^{*}p < 0.05$, $^{**}p < 0.01$,$^{***}p < 0.001$.

A simple regression of the fraction choosing a particular major on the fraction of workers with that major in their home county is shown in Table 2.1. The correlation between local labor market conditions and major choices is significant even after controlling for major and county fixed effects.

Clearly, these results are partially explained by greater demand for particular occupations in specific home counties. Combined with the fact that it is costly to move between counties, this would be sufficient to drive the positive relationship documented in this section. However, such employment specific outcomes would seem to be especially important for technical occupations which require STEM degrees rather than in occupations that are more general purpose (business and social science) and services which are broadly needed (education and medicine). We see exactly the opposite trends in Figure 2.2. These results are in fact much more naturally explained by informational frictions.

2.2.3 Inter-County Migration has Slowed

The local bias in education wouldn't be problematic if skills were redistributed across locations. This would reduce the local bias in skill composition. However, overall rates of inter-county migration have been slowing substantially in recent years as documented in Molloy *et al.* (2011). There is some distinction however between skilled and unskilled migration. Using Current Population Survey data from the years 1985-2019, I show trends

in the migration rates for people 25-55. I subset to this age range to approximate a time after receiving a degree and before retirement when migration ceases to have an impact on labor market composition.

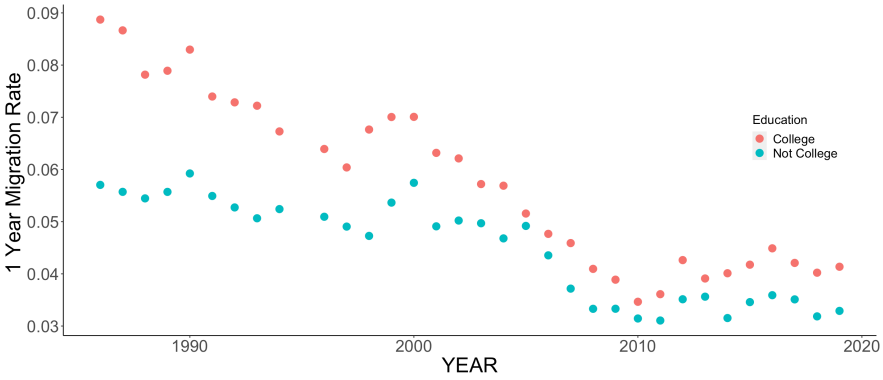


Figure 2.3: *CPS Intercounty Migration Rates*

Note: This figure is based on CPS data from 1985-2019. The population is subset to ages 25-55 with education levels based on whether or not the respondent had a college degree.

The downward trend shown in Figure 2.3 is dramatic for both college and non-college educated workers. As of 2010, the United States seems to have reached a plateau of consistently lower migration rates although it is still the case that those with a college degree are more mobile. Stange *et al.* (2022) look at migration patterns for recent graduates more closely and find that according to publicly available LinkedIn data, 50% of recent college grads stay in the MSA in which the institution they attended was located. This is especially true for less selective universities. The low migration rates post graduation are further exacerbated by the fact that the median 4-year public university student travels 18 miles to attend their institution according to the American Council on Education (Hillman and Weichman (2016)).

In summary, I have shown that there exists a large variation in the returns to skill both at the macro level (college vs. not) and at the micro level (college major). This variance explains findings that students do a poor job of predicting expected returns to specific majors. In the face of this uncertainty there seems to be a tendency to pick a major which is well

represented in a students home labor market. The pile-on effect here is then exacerbated by decreased migration rates. In the next section, I develop a structural spatial model which captures these facts in order to show how local bias in skill acquisition can affect the geographic distribution of skills leading to persistent regional inequality.

2.3 Structural Spatial Model

The objective of this section is to build off canonical structural spatial models while including the impact of educational choice allowing for a positive externality associated with local labor markets. We will refer to the agents in this economy as students because the key decisions take place before completing their studies. Students choose type and level of education, $m \in \{0, 1, ..., M\}$, where there are M majors and $m = 0$ refers to forgoing a college degree. Attaining a particular education results in an ability level, $\alpha_{c_o m}$, which depends on the location of origin, c_0 and educational choice. Each student can choose to move to any city, $c \in C$. In order to move from origin city c_0 to destination c, a moving cost of $d_{c_0 c}$ must be paid where $d_{c_0 c_0} = 0$. Once in that location they inelastically supply labor and receive the prevailing wage which is a function of their ability.

2.3.1 Preferences:

As in Davis and Dingel (2019), each student will consume λ units of a local non-tradeable good, a tradeable good and one unit of housing. They also have idiosyncratic tastes over major-city pairs denoted by $z_{c_0 mc}$, which can in principle depend on the origin. For tractability, I assume that these shocks are distributed independently Frechet across major-destination pairs such that

$$F_{c_0 mc}(z) = e^{-T_{c_0 mc} z^{-\theta}}$$

Absolute preference for particular major-destination pairs varies by home location c_0 through $T_{c_0 mc}$, and the dispersion of preferences is inversely related to θ. An important assumption for this analysis is that the idiosyncratic component of preferences is specific to educated

workers and is realized after the decision of whether or not to get a college degree. This captures the fact that shocks which affect preferences and costs take place during university before having committed to particular majors or destinations. Although this is a significant assumption, it greatly simplifies the estimation procedure in Section 2.4. This assumption is summarized in the following assumption:

Assumption 1 *The idiosyncratic taste component of preferences is only significant for skilled workers.*

In order to account for the risk that students face in deciding on what skills to invest in, students have CARA utility functions. This formulation allows for tractable results with the Frechet preference shock. Combined, this implies that student's indirect utility is given by

$$V_i(c_0, m, c) = -\frac{1}{z_{c_0 mc,i}} \exp\left(-w_i(\alpha_{c_0 m}, c) + t_{c_0 m} + \lambda p_{nc} + p_{hc} - a_c + d_{c_0 c}\right) \tag{2.1}$$

which depends on the prevailing wage for workers with ability $\alpha_{c_0 m}$ in city c, $w(\alpha_{c_0 m} c)$, as well as the price of nontradeables, p_{nc}, housing, p_{hc}, a local amenity a_c and the cost of education,$t_{c_0 m}$.

2.3.2 Learning:

This section introduces the main departure of the model from canonical formulations. There are many reasons why we might expect students to be affected by their local labor market. There may be direct effects along the lines of Marshall's quote in the introduction: being around people with a particular education might endow you with the necessary ability to succeed. This would effectively reduce the cost of obtaining particular skills. It may also be the case that people surrounded by certain occupations grow to like them better. This would add an amenity term for particular occupations. While both of these are reasonable explanations, I will focus on a third which is that being surrounded by people with particular skills reduces the uncertainty about obtaining those skills.

The reduction in uncertainty is in line with findings in the education literature which highlights the lack of information students have about future earnings and their risk aversion

(Altonji *et al.* (2016), Andrews *et al.* (2022), Altonji *et al.* (2012)). Specifically, Wiswall and Zafar (2015) find that students exhibit a high level of risk aversion which significantly affects their major choice. The source of uncertainty that is revealed by local labor markets may be with regards to the returns to ability or to the cost of acquiring skills. This model will be framed in terms of uncertainty about the wage although it is isomorphic to alternative interpretations. The general form that this derivation simplifies to is both a gift and a curse in that one needs to either be agnostic about the underlying reason for the effect or must lean on the identification strategy in order to pick out a specific component. For more discussion about the learning channel and some reduced form evidence see Appendix Section B.1.

2.3.3 Signal Extraction Model:

From the standpoint of the individual, the wage is taken as given although there is uncertainty as to what the student's future wage will be when they enter the labor market. I assume that students consider the wage to be a normal random variable which depends on your ability (origin-major choice) and the destination specific production technology. The wage they are trying to learn about takes the form:

$$w(c_0, m, c) = \mu_{\alpha_{c_0 m c}} + \xi_{c_0 m c}$$

where $\mu_{\alpha_{c_0 m, c}}$ is the mean wage and $\boldsymbol{\xi}$ is the variance which follows a multi-variate normal distribution, $\boldsymbol{\xi} \sim N(0, \boldsymbol{\Sigma})$. The variance for a given origin-major-destination wage is given by $\sigma^2_{c_0 m c}$. The common prior about the distribution of mean wages is given by $\mathbf{w_0} \sim N(\bar{\mathbf{}}_{\mathbf{0}}, \boldsymbol{\Sigma})$. Here $\mathbf{w_0}$ is a C x C x M vector of wages, which includes all possible combinations of majors across origin and destination cities. When making their education/location decision an individual from c_0 will only use information about the relevant 1 x C x M components of the wage vector. Similarly the correlation structure of wages across locations is given by $\boldsymbol{\Sigma}$, which is known.

Learning takes place by observing a discrete number, κ, of observations drawn randomly

from the origin city. I'll assume that there is migration across destinations such that the wage observed depends on migration patterns: if more people from Boston move to San Francisco, people in Boston become more aware of what wage they would be expected to receive in San Francisco. Specifically, let $\boldsymbol{M}$ be the migration matrix which defines the fraction of individuals from the previous generation that moved across cities where $M_{cc'}$ is the fraction of individuals who moved from c to c'. The vector of the number of observations that someone from origin city c_0 observes across destinations is given by $\kappa \boldsymbol{M_{c_0}}$. We can define the fraction of individuals from c_0 with major m to be $\delta_{c_0 m}$. For each location they observe $\kappa M_{c_0 c} \delta_{c_0 m}$ observations of the origin-major-destination specific wage signal:

$$w(c_0, m, c) = \mu_{\alpha_{c_0 m} c} + \xi_{c_0 m c} + \hat{\xi}_{c_0 m c}$$

where $\hat{\xi}_{0mc}$ is the signal noise distributed $N(0, \sigma^2_{\hat{\xi}, c_0 m c})$. The diagonal of the signal precision matrix is given by

$$P_{c_0 m c} = \left(\sigma^2_{c_0 m c} + \sigma^2_{\hat{\xi}, c_0 m c}\right)^{-1} \kappa M_{c_0 c} \delta_{c_0 m}$$

Updating results in the following:

$$\boldsymbol{w} \sim N\left((\boldsymbol{\Sigma}^{-1} + \boldsymbol{P})^{-1}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu_0} + \boldsymbol{P}\boldsymbol{\mu}), (\boldsymbol{\Sigma}^{-1} + \boldsymbol{P})^{-1}\right)$$

The vector of mean wage signals is given by $\boldsymbol{\mu}$. We'll denote the posterior distribution to be given by $N(\boldsymbol{\mu^p}, \boldsymbol{\Sigma^p})$

2.3.4 Expected Indirect Utility:

Now that we have put structure on the uncertainty faced by students in this model, we can evaluate expected indirect utility. In this model, the impact of local labor markets on education, beyond direct effects captured by the wage, come through agents' risk aversion. An individual will penalize the wage as a function of the number of signals of each major-destination combination that they observe. Taking expectations of Equation 2.1 conditional on the Frechet shock results in:

$$E_{c_0}[V_i(c_0, m, c)] = \tag{2.2}$$
$$-\frac{1}{z_{c_0mc,i}} \exp\left(-\mu^p_{c_0mc} + \frac{1}{2}\left(\sigma^2_{c_0mc} + \Sigma^p_{c_omc}\right) + t_{c_0m} + \lambda p_{n,c} + p_{h,c} - a_c + d_{c_0c}\right)$$

We now assume that students have diffuse priors about the wage such that their only source of information is through wage signals that they receive.

Assumption 2 *: Students have diffuse priors such that their posterior beliefs are entirely dependent on the signalling structure of their local labor market.*

Then we have that

$$\Sigma^p_{c_omc} = \frac{\left(\sigma^2_{\xi,c_0mc} + \sigma^2_{\hat{\xi},c_0mc}\right)}{\kappa M_{c_0c}\delta_m} \tag{2.3}$$

This makes clear that the indirect utility associated with a particular origin-major-destination is an increasing function of the number of workers in their local labor market with the relevant major and the level of migration to a particular destination.

2.3.5 Goods Production

I abstract from the production side of the economy to consider only labor demand. We have already seen that wages are given by

$$w(c_0, m, c) = \mu_{\alpha_{c_0mc}} + \xi_{c_0mc}$$

In this section I bring in the traditional productivity effects of agglomeration. There is a strict difference between $m = 0$ and all other major choices. Specifically, it is assumed that students who choose to forego a college degree will be employed in the non-tradable sector. Productivity in the non-tradable sector is the same across all origins and destinations. This sector can be thought of as the low-skill services sector. Because productivity is the same everywhere, it is easier to learn and thus, I'll assume that the wage for this sector is known with certainty. For all other workers, productivity takes the form of a destination specific productivity, ρ_c, a normally distributed productivity draw $h_{\alpha_{c_0mc}}$ and an agglomeration

component taken to be the average of all individual productivity levels in the city. This is a simplified version of the framework in Davis and Dingel (2019). Combined, the mean wage for college graduates is given by

$$w_{c_0mc} = \rho_c h_{\alpha_{c_0mc}} H_{cm}^{\gamma_\alpha}$$
$$H_{cm} = \frac{1}{L_c} \int_{\{i|c_i=c, m_i=m\}} h(\alpha_i, c) di$$

where for individual i from origin c_0 with major m, $h(\alpha_i, c) = h_{\alpha_{c_0mc}}$ and L_c is the total population in destination c. This formulation highlights that students are learning about the overall city productivity vector $\boldsymbol{\rho}$, the level of agglomeration $\boldsymbol{H}$, and an origin-major specific wage component, $\boldsymbol{h}$. This last piece can be thought of as a matching function between the specific abilities produced by a certain origin-major pair and the skills demanded across cities.

Altogether we have that the mean wage is given by

$$w_{c_0mc} = \begin{cases} \rho_c h_{\alpha_{c_0mc}} H_{cm}^{\gamma_\alpha} & \text{if } m \neq 0 \\ p_{n,c} & \text{if } m = 0 \end{cases} \tag{2.4}$$

A key feature of this model will be how the productivity effects of agglomeration which come about by concentrating skills interacts with the signaling effect.

2.3.6 Housing

The price of housing is given by a constant elasticity price equation which depends on the population in the city, L_c.

$$p_{h,c} = \kappa L_c^{\gamma_h} \tag{2.5}$$

2.3.7 Solving the Model

The solution to this model is a Nash Equilibrium, where all students make the optimal major-destination choice accurately anticipating the wage and labor markets clear. In order

to simplify notation, I'll denote the effective wage ω_{c_0mc} as the value of the wage taking into account uncertainty net of education cost:

$$\omega_{c_0mc} \equiv \mathbf{1}_{m>0}\left(\mu^p_{c_0mc} - \frac{1}{2}\left(\sigma^2_\xi + \Sigma^p_{c_0mc}\right) - t_{c_0m}\right) + \mathbf{1}_{m=0}\left(p_{nc}\right) \tag{2.6}$$

This allows us to write the major-destination optimization problem as:

$$\max_{m\in\{0,M\},c\in C} E_{c_0}[V_i(c_0,m,c)] = \max_{m\in\{0,M\},c\in C} -\frac{1}{z_{c_0mc}(i)}\exp\left(-\omega_{c_0mc} + \lambda p_{n,c} + p_{h,c} - a_c + d_{c_0c}\right)$$

Taking advantage of the properties of the Frechet distribution, we can then calculate the probability that a student with a college degree from c_0 chooses major m and city c as

$$P(c,m|c_0) = \frac{T_{c_0mc}\exp\left[\theta\left(\omega_{c_0mc} - \lambda p_{n,c} - p_{h,c} + a_c - d_{c_0c}\right)\right]}{\sum_{m'\in\{0,M\}}\sum_{c'\in C} T_{c_0m'c'}\exp\left[\theta\left(\omega_{c_0m'c'} - \lambda p_{n,c'} - p_{h,c'} + a'_c - d_{c_0c'}\right)\right]} \tag{2.7}$$

For unskilled labor it must be the case that indirect utility is equalized across locations c for someone from c_0. This implies

$$(1-\lambda)p_{nc} + a_c - p_{h,c} - d_{c_0c} = \overline{v}_{c_0} \quad \forall \quad c \in C \tag{2.8}$$

Because all students are ex ante identical we also have that there must be an indifference condition between getting a college degree and not. This condition is analogous to the price index in Eaton and Kortum (2002) so that the value of college is proportional to the "market potential" with the appropriate elasticity.

$$\overline{v}_{c_0} = \Xi\Phi^{1/\theta}_{c_0} \tag{2.9}$$

where

$$\Phi_{c_0} = \sum_{m'\in\{0,M\}}\sum_{c'\in C} T_{c_0m'c'}\exp\left[\theta\left(\omega_{c_0m'c'} - \lambda p_{n,c'} - p_{h,c'} + a'_c - d_{c_0c'}\right)\right]$$

From market clearing we have that

$$L = \sum_c L_c \tag{2.10}$$

and

$$\delta_{cm} L_c = \sum_{k \in C} [L_k P(c, m|k)] + \underline{L}_{cm} \tag{2.11}$$

where $\underline{L}_{cm}$ is the initial number of workers in city c with major m. This equation simply states that the number of people who end up with major m in city c must be equal to the number who decide to get that major and move to that city across all other location plus the initial endowment.

Additionally, the market for non-tradable production must clear such that

$$\delta_{cm=0} L_c = \lambda L_c \quad \forall c$$

This condition will pin down the price of the non-tradable good and therefore, the low skilled wage across locations.

2.3.8 Comparing to the Existing Literature

In order to compare this to previous work on agglomeration, I can simplify the model to focus on the choice of whether or not to obtain a college degree, $m \in \{0, 1\}$. The key point of emphasis is on the determinants of the effective wage, $\omega_{c_0 mc}$. We can rewrite Equation 2.6 substituting in for the determinants of the true wage in a setting where Assumption 1 holds:

$$\omega_{c_0 mc} \equiv \mathbf{1}_{m=1} \left(\rho_c h_{c_0 c} \delta_c^{\gamma_\alpha} - \frac{1}{2} \left(\sigma_{\tilde{\xi}}^2 + \frac{\left(\sigma_{\tilde{\xi}, c_0 c}^2 + \sigma_{\tilde{\xi}, c_0 c}^2 \right)}{\kappa M_{c_0 c} \delta_{c_0}} \right) - t_{c_0} \right) + \mathbf{1}_{m=0} \left(p_{n,c} \right) \tag{2.12}$$

Note that I have used the fact that it is without loss to consider $H_c = \delta_c$. Now we can see that the effective wage for college educated workers depends on the concentration of skill in the **destination** through a positive productivity effect, but it also depends on the concentration of skill in the **origin** through a signaling effect. The question of place-based policy from the perspective of optimal skill allocation then hinges crucially on the relative value of these two effects.

2.4 Model Simplification and Estimation

The calibration I present in this paper is meant to be a first pass in understanding the impact of labor market signaling on skill acquisition. I leverage migration data from the ACS as well as local wages and housing prices in order to back the key structural parameters in the model. This will allow me to perform a counterfactual where we begin from an equal distribution of skills across locations. In order for the calibration to be consistent with recent work such as Diamond (2016), I focus on the two skill case as in Section 2.3.8.

2.4.1 Data

The primary moment that I will target in this analysis is given by Equation 2.7 and represents migration shares across origin-destination pairs. With this in mind, the primary data used in this calibration is decennial census data for the years 1980, 1990 and 2000 accessed through IPUMS. In these years, a 5% sample of the population was asked to state where they lived 5 years ago. This allows me to back out the migration matrix which forms the back bone of the calibration.

I also use the education data from the census, which is limited to coarse measures of educational attainment and does not distinguish by major. Although I could proxy for choice of major by industry and occupation, this would lead to substantial noise. The ACS has data on major choice but has more limited migration information and is only available from 2009 onwards. This data also includes household rental rate which is used as a proxy for local housing prices. I also deflate all wages and prices using national consumer price index.

2.4.2 Non-Tradable Component and Amenities

The first step in the calibration is to solve for λ which represents the amount of non-tradable consumed. The model suggests there are two distinct ways to calculate this. The first method is to realize that the indifference condition shown in Equation 2.8 implies a specific relationship between the local unskilled wage, p_{nc} and the price of housing, p_{hc}, which

depends on λ. In order to simplify this analysis I'll make the additional assumption that the cost of moving is equal across destinations. Clearly, this will limit the models ability to capture the variation in migration patterns. However, this aides the step-by-step estimation of the model and will be validated in part by looking at the city-specific amenity levels that are dependent on this assumption.

Assumption 3 *There are no migration costs.*

This assumption leads to a simplified version of Equation 2.8:

$$(1-\lambda)p_{nc} + a_c - p_{h,c} = \overline{v} \quad \forall \quad c \in C \tag{2.13}$$

Rearranging, this result yields an econometric means of evaluating λ by regressing the price of housing on the price of non-tradables which through the lens of the model is the unskilled wage. Using the Census data described in Section 2.4.1, I estimate the following specification:

$$p_{hct} = (1-\lambda)p_{nct} + a_c + \epsilon_{ct} \tag{2.14}$$

Note that I've include a time subscript t to emphasize that observations are at the MSA-Year level. The value of p_{nc} is taken to be the average wage of workers in an MSA that do not have a college degree. Note that amenity term a_c is a county level fixed effect. I assume that λ and a_c do not change across the time period. The result of this analysis is a λ of 0.703. The regression also produces amenity values. In Table 2.2, I list the top and bottom 5 MSAs in terms of their amenity value that comes out of the regression in Equation 2.14. These MSAs provide some reassurance as some of the warmer and nicer areas in the country appear in the top 5 while some of the colder and poorer areas for this time period appear in the bottom 5.

An alternative means of calculating λ, is noting that in aggregate it must be that $\lambda = \frac{L_{m=0}}{L}$. With this alternative formulation I get $\lambda = .729$. The list of high and low amenity MSAs is roughly the same with Honolulu, HI replacing State College, PA in the top 5. These

Table 2.2: *High and Low Amenity MSAs*

Top 5		Bottom 5	
MSA	Amenity Value	MSA	Ameinty Value
Santa Barbara, CA	$4560	Flint, Michigan	-$3170
Washington D.C.	$4310	East Chicago, IN	-$2520
Gainseville, FL	$4100	Detroit, MI	-$2060
San Jose, CA	$4010	Sheboygan, WI	-$1710
State College, PA	$3890	Green Bay, WI	-$1670

Note: Results are based on a regression of of annualized rents on low skilled wages across MSAs from Census data for the years 1980, 1990 and 2000. Amenity values are on anualized basis.

differences do not fundamentally affect the analysis and going forward I will lean on the first specification with $\lambda = .703$. It is important to note that both of these estimates and the subsequent amenity values are heavily dependent on the choices made in our structural model. Future work may perform a more robust analysis as in Diamond (2016). However, it is beyond the scope of this analysis.

2.4.3 Signaling Impact

I can evaluate the impact of the fraction of college worker's in a students local labor market on their educational choice by simplifying and estimating Equation 2.7. To simplify this equation it is useful to take logs and simplify for the binary education decision resulting in

$$\ln P(c|c_0) = \ln T_{c_0c} + \theta\left(\omega_{c_0c} - \lambda p_{nc} - p_{hc} + a_c - d_{c_0c}\right) - \ln(\Phi_{c_0})$$

where Φ_{c_0} is the college educated market potential for location c_0. I then build up our estimating equation by subtracting the probability of staying in the home location resulting in

$$\ln\left(\frac{P(c|c_0)}{P(c_0|c_0)}\right) = \ln\left(\frac{T_{c_0c}}{T_{c_0c_0}}\right) + \theta\left((\omega_{c_0c} - \omega_{c_0c_0}) - \lambda(p_{nc} - p_{nc_0}) - (p_{hc} - p_{hc_0}) + (a_c - a_{c_0})\right)$$

I further make three simplifications to get to our key estimating equation. The first is

to simplify the absolute advantage. I am currently allowing for there to be an absolute advantage in taste across specific origin-destination pairs. While this would help capture recurring flows as in Schubert (2021), limitations in the data make estimating these parameters challenging. Alternative approaches would be to allow for there to be an advantage for the origin location, but instead I will simply assume that there is no absolute advantage in taste for location after the education decision. Thus, I will make the simplification that

$$T_{c_0 c} = 1 \quad \forall \quad c_0, c \in C.$$

The second simplification is related to the effective wage. From Equation 2.6 and substituting in Equation 2.3, the effective wage for students with a college degree is

$$\omega_{c_0 c} = \mu_c^p - \frac{1}{2}\left(\sigma_{\tilde{\xi}}^2 + \frac{\left(\sigma_{\tilde{\xi},c_0 mc}^2 + \sigma_{\hat{\tilde{\xi}},c_0 mc}^2\right)}{\kappa M_{c_0 c}\delta_{c_0}}\right) - t_{c_0}$$

Consider the migration matrix, $M_{c_0 c}$. Most individuals stay in their local area whereas a much smaller number migrates to a particular destination. This implies that the major impact on the signal strength is whether the student is considering the origin location vs. an alternative destination. Furthermore, I'll allow for the potential for an elasticity other than 1 with respect to the fraction of college workers in the origin. Together this implies

$$\omega_{c_0 c} = \mu_c^p - \tilde{\sigma}_{\tilde{\xi}}^2 - \tilde{\zeta}\delta^{-\gamma}\left(1 + \tau \mathbf{1}_{c \neq c_0}\right) - t_{c_0}$$

where

$$\tilde{\sigma}_{\tilde{\xi}}^2 = \frac{1}{2}\sigma_{\tilde{\xi}}^2 \quad \text{and} \quad \tilde{\zeta} = \frac{1}{2}\frac{\left(\sigma_{\tilde{\xi},c_0 mc}^2 + \sigma_{\hat{\tilde{\xi}},c_0 mc}^2\right)}{\kappa}$$

and τ represents the lower signal intensity associated with cities that are not the origin.

The last assumption is to set $\theta = 1$, which reduces the estimating equation to something that is more akin to a standard discrete choice model with Type 1 Extreme Value shocks. Altogether, we can substitute in the effective wage and make the stated simplifications to get the moment that we will be estimating in the data:

$$\ln\left(\frac{P(c|c_0)}{P(c_0|c_0)}\right) = (w_c - w_{c_0}) - \zeta\delta^{-\gamma} - \lambda(p_{nc} - p_{nc_0}) - (p_{hc} - p_{hc_0}) + (a_c - a_{c_0}) \quad (2.15)$$

where $\zeta = \tilde{\zeta}\tau$. On the right hand side of Equation 2.15, everything is known except for ζ and γ. This means we can rearrange the above equation to get our estimating equation. I further allow ζ to vary by year as this will help soak up potential trends over time, so I estimate

$$\ln(Res) = \ln(\zeta_t) - \gamma \ln \delta$$

where $Res = \ln\left(\frac{P(c_0|c_0)}{P(c|c_0)}\right) - (w_{c_0} - w_c) + \lambda(p_{nc_0} - p_{nc}) + (p_{hc_0} - p_{hc}) - (a_{c_0} - a_c)$. The results shown in Table 2.3 has the expected sign of γ and is statistically significant with standard errors clustered at the origin level.

Table 2.3: *Estimates of Key Model Parameters Based on Migration Probabilities*

	Estimate	s.d.	p-value
γ	0.61	0.09	0.00
ζ_{1980}	7.26	0.16	0.00
ζ_{1990}	7.59	0.14	0.00
ζ_{2000}	8.03	0.12	0.00

I also convert this estimate into the dollar impact comparing originating in a county with a twentieth percentile vs. eightieth percentile college fraction in Table 2.4. In columns (a) and (b) show the variation in college fraction across each census year investigated. Column (c) shows the impact of local college fraction through the informational channel in dollar terms. Shifting from a low college to high college origin increases the perceived value of college by $1200 in 1980. We can then compare this to the standard deviation of the value across all locations which includes the impact of wages, prices and amenities. We can see that the impact of local labor markets through channels other than the wage can change the perceived utility by roughly 1/5 of the amount of all factors combined across locations.

Table 2.4: *Impact of Local College Fraction in Dollar Terms*

Year	(a) δ_{P20}	(b) δ_{P80}	(c) Impact	(d) SD
1980	0.16	0.26	\$1200	\$6300
1990	0.18	0.31	\$1500	\$7100
2000	0.21	0.35	\$2100	\$10900

2.4.4 Agglomeration Elasticity

The last value to estimate in order to perform our simple empirical exercise is the agglomeration elasticity. For this I again follow the simplification outlined in Section 2.3.8. I also simplify the origin-destination component such that $h_{c_0c} = h_c$. In order to determine the elasticity, I simply perform an OLS regression of skilled wage on the fraction of skilled workers in each destination.

$$\ln(w_c) = \ln(\rho_c h_c) + \gamma_\alpha \ln \delta_c$$

The results are shown in Table 2.5. Here we get a positive and significant value of the agglomeration coefficient. It is significantly smaller than the value obtained for the signaling elasticity. This implies a much larger impact of college fraction on wages when moving from the twentieth to the eightieth percentile as observed in Table 2.6.

Table 2.5: *Estimating Agglomeration Effect of Destination College Fraction*

	estimate	s.d.	p-value
γ_α	0.22	0.06	0.00

This analysis takes the destination MSA across decadal census year as the unit of observation. This regression includes both MSA and Year fixed effects for the years 1980, 1990, 2000.

2.5 Experiment

In order to get a sense for how the local labor market affects student outcomes, I evaluate the partial equilibrium impact of a more equal distribution of college workers on student

Table 2.6: *Impact of Destination College Fraction on Wage*

Year	(a) δ_{P20}	(b) δ_{P80}	(c) Impact	(d) SD
1980	0.16	0.26	\$7900	\$6300
1990	0.18	0.31	\$9100	\$7100
2000	0.21	0.35	\$10300	\$10900

expected utility. This simplified model is not suited for a full structural analysis, but it can show us where some of the impacts of local labor markets are coming from and how they compare to agglomeration effects. The key parameter of interest is the market potential, Φ, which in equilibrium will be proportional to the expected indirect utility for both college and non-college students through Equation 2.9.

I consider what would happen if we distribute the total number of college workers across all MSAs such that $\delta_c = \delta \quad \forall c \in C$. The key question is how does the market potential change in response to the change in college fraction across all locations. I can further break this down to include or exclude the component due to local labor market signaling. In our simplified model we have

$$\Phi_{c_0} = \sum_{c \in C} \exp\left[\rho_c h_c \delta_c^{\gamma_\alpha} - \tilde{\sigma}_\epsilon^2 - \tilde{\zeta}\delta_c^{-\gamma}(1 + \tau_{c \neq c_0}) - \lambda p_{n,c'} - p_{h,c'} + a'_c\right]$$

I will further simplify this by noting that the variance term $\tilde{\sigma}_\epsilon^2$ only adds a constant of proportionality. Further I will assume that the origin mean wage is known with certainty such that the only uncertainty is the wage in alternative destinations. Mathematically this implies that the signaling uncertainty reduces to $\zeta\delta_c^{-\gamma}$. Lastly, the exponential causes issues in the partial equilibrium because it leads to explosive results. Therefore, our key parameter of interest will be an analogous term where I eliminate the exponential. This parameterization will allow me to explore the impact of the change in college fraction heuristically:

$$\tilde{\Phi}_{c_0} = \sum_{c \in C} \left[\rho_c h_c \delta_c^{\gamma_\alpha} - \zeta_{c \neq c_0}\delta_c^{-\gamma} - \lambda p_{n,c'} - p_{h,c'} + a'_c\right] \tag{2.16}$$

In Figure 2.4, I've plotted both the change in welfare associated with agglomeration alone as well as accounting for the local labor market signal. Because we have simplified the model to be a flat world without differential moving costs across destinations, the impact of the agglomeration component is the same across all locations independent of the original college fraction in that location.

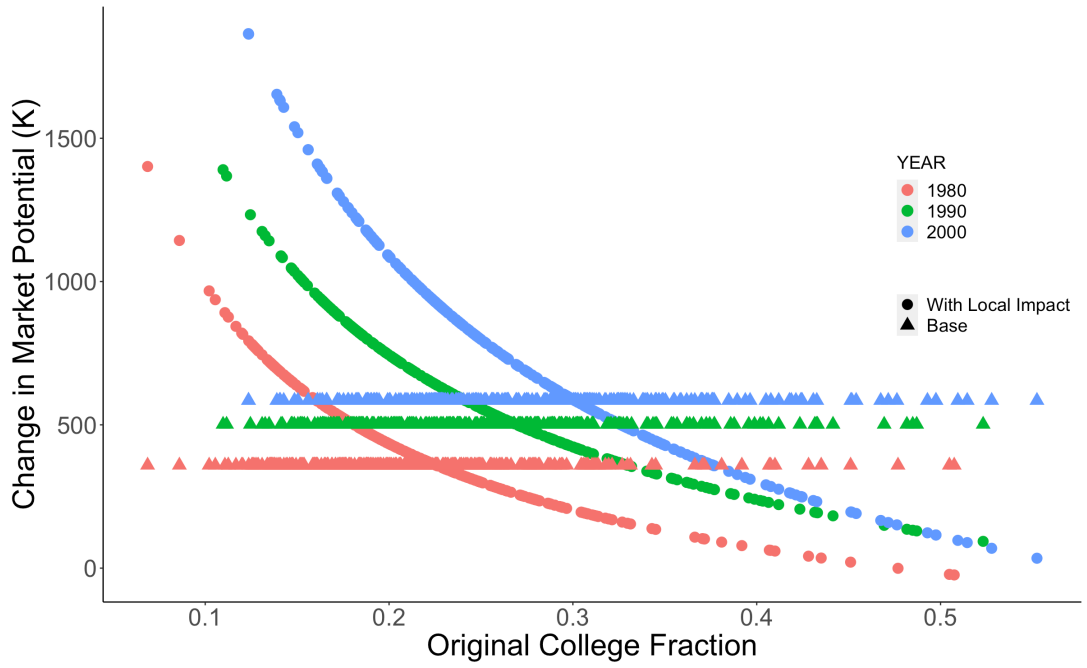


Figure 2.4: *Impact of Redistributing College Workers on Welfare*

Note: The change in a simplified measure of market potential are shown across different decades. This plot shows the change including the local labor market impact of signaling as well as without it. The baseline positive effect is due solely to the benefits of agglomeration from redistribution.

The effect of redistributing workers is positive here due to the concavity of the agglomeration function. The impact of skill redistribution across locations shows up in the impact on the risk term associated with signaling. Here we see that locations with low initial college fractions benefit substantially from redistributing skilled workers. These gains offset the losses experienced by students in initially high-skilled locations. Again, this is due to the convexity in the cost associated with a weak local labor market signal. These relative gains are also increasing across decades as the overall skill level in the population is increases.

2.6 Conclusion

The main finding of this paper is that allowing for the composition of local labor markets to affect educational choice through signaling the returns to particular skills or similar channels can lead to significant heterogeneity in terms of the efficacy of place-based policies. This impact is an extension of the findings that there is substantial misinformation that students have with regards to the return to skill. The importance of signaling specifically aligns with evidence on the impact of role models and other interventions which can significantly affect major choice. I provide evidence in this paper that local labor markets tend to influence major choice beyond the affect on local wages.

In light of these findings it is unsurprising that migration patterns have slowed as shown in Figure 2.2. Without major shake-ups in terms of labor market composition, we would expect the accumulation and concentration of particular skills in specific locations through a bias towards local labor market skills. This divergence would then lead to more mismatch between skills acquired in a particular location compared to those demanded in other locations.

The objective of this paper is to spur further research into the spatial determinants of educational choice. Externalities which lead to biases in terms of skill acquisition related to local conditions (as is the case with signaling) will lead to spatial divergence in terms of outcomes and will reduce aggregate productivity. Policy recommendations which suggest the potential for concentrating/diversifying skills across space should makes sure to consider the impact that this can have on future generations of workers. Specifically, allowing for the spatial concentration of in-demand skills may substantially reduce the likelihood that students in less skill intensive regions invest in high return skills. It should be a policy goal to internalize these signaling externalities.

These concerns should also be investigated to see how they interact with the availability of educational resources across space. Given that students tend to stay near home when pursuing their degrees the presence of educational deserts can be particularly devastating and add to the limited set of opportunities which students may consider. Beyond this,

the resources at the major level may differ across space and may themselves be biased towards local labor markets. This may be advantageous when considering the creation of technology hubs but may make matching local skill acquisition to aggregate skill demand more challenging.

A lot of attention has been focused on the evolution of the skill premium, but less research has been focused on how to ensure more people are able to benefit from the increasing skill premium. In this paper I propose an important explanation which has to do with the mismatch between local and global skill demands. However, there are many more reasons why we seem to fear the widening skill premium rather than embracing the ability for people to shoot up the ladder by taking advantage of the high wages associated with skill upgrading. I hope that more energy will be spent discussing this aspect of inequality.

Chapter 3

Bilateral Contracts and Social Welfare[1]

3.1 Introduction

Contract law is composed of “default" and “immutable" rules. Default rules can be contracted around, but will be enforced if the contracting parties do not agree to an alternative. Meanwhile, immutable rules cannot be modified. They will be enforced even if the parties attempt to contract around them. The literature generally agrees that the law uses immutable rules to constrain contracting when parties within or outside the contract cannot adequately protect themselves. That is, immutable rules mitigate “internalities" and externalities.[2]

The role of default rules is less well understood and is a subject of intense debate in legal theory. A common view is that default rules should supply the contract terms that parties “would have wanted" if they had full information and could bargain costlessly.[3]

This paper provides a theoretical analysis of the role of immutable and default rules. It highlights that immutable rules manage the tradeoff between promoting efficiency and

[1] Co-authored with Zoë Hitzig

[2] See Ayres and Gertner (1989) for a discussion of the legal literature on immutable and default rules.

[3] A few expressions of the “would have wanted" view of default rules, as quoted in Ayres and Gertner (1989): Easterbrook and Fischel (1989) write that the default rule should be “the term that the parties would have selected with full information and costless contracting;" Posner (1986) writes that the default rule should supply “the contract that most well-informed persons would have adopted if they were to bargain about the matter;" Goetz and Scott (1983) write, “the preformulated rules supplied by the state should mimic the agreements contracting parties would reach were they costlessly to bargain out each detail of the transaction."

curbing externalities, while default rules ensure particular distributions of surplus across unverifiable states of the world.

There are many settings in which governments have concerns about the distribution of surplus between contracting parties, which we will refer to as “equity" concerns. In incomplete contracting environments, default rules can reduce the risk of non-contractible states of the world by ensuring a desirable distribution of surplus in those states. In other settings, the law may seek to ensure a particular distribution of surplus for fairness reasons. For example, the laws governing marital contracts in many U.S. states enforce default rules that stipulate an “equitable division" of assets in the event of divorce. In both of these settings, the default rules can be seen as achieving distributional objectives while immutable rules internalize externalities. More generally, we argue that the tradeoffs between efficiency, equity and externalities affect the optimal policy whenever the state seeks to influence the outcome of bilateral contracts.

In our model, a social planner or government (whom we call the “regulator") designs the contracting environment in order to maximize social welfare. How the contracting parties value features of the contract is information (the “state of the world") that is observable (to the agents) but not verifiable (by the regulator). The regulator uses a decentralized indirect mechanism, inspired by default and immutable rules, which we call *default delegation*. This mechanism involves choosing a *default* which serves as a disagreement point in the parties' negotiations, and also has a *delegation* aspect, in that agents are able to bargain from the default to their preferred contract within a restricted set of enforceable contracts.

Default delegation thus joins ideas about renegotiation in incomplete contracts to theories of delegation. Our model is close in spirit to the model of renegotiation design for incomplete contracts in Aghion *et al.* (1994), but differs in that we introduce a regulator whose preferences may not align with the contracting parties' preferences. We differ from classical models of delegation (Holmström, 1977, 1984; Alonso and Matouschek, 2008) in that a principal chooses a delegation set for two agents who efficiently bargain instead of delegating to a single agent.

The first major contribution of this paper is to provide mechanism design foundations for the use of immutable and default rules in contract law. With this in mind, we start from a general analysis of implementable outcomes based on the model of implementation with renegotiation in Maskin and Moore (1999). We then characterize the conditions under which default delegation achieves first-best social welfare, which depend on the degree to which efficiency, equity and externalities are weighted in the regulators' social welfare function. The result of this analysis is a mechanism design foundation for default and immutable rules: default rules ensure particular distributions of surplus while immutable rules curb externalities. In laying bare the tradeoffs inherent in the choice of default and immutable rules, our analysis illustrates the broader appeal and applicability of default delegation mechanisms.

The second major contribution of this paper is to provide comparative statics for the optimal default delegation policy. This analysis illustrates how the optimal policy depends on the social welfare function as well as other primitives such as bargaining parameters and the regulator's beliefs with regard to the true state of the world. Furthermore, we apply our results to the regulation of app-based platform work, illustrating how our results deliver insights into the design of regulation and public policy.

The regulation of app-based platform work provides a useful case study because regulators seeking to pass laws about platform work have stated objectives including efficiency, equity **and** externalities. For example, Secretary of Labor Marty Walsh, who has been vocal about rethinking the classification of platform workers, spoke to the need to balance efficiency *and equity* in early 2021: "These companies are making profits and revenue and I'm not (going to) begrudge anyone for that... But we also want to make sure that success trickles down to the worker."[4] California's Assembly Bill 5, which attempted to reclassify platform workers as employees in 2019, cited the *externalities* that may accompany misclassification. The court noted the potential harm borne by taxpayers due to "the loss to the state of

[4] https://www.washingtonpost.com/business/2021/04/29/labor-walsh-gig-workers-employees/

needed revenue from companies that use misclassification to avoid obligations such as payment of payroll taxes, payment of premiums for workers' compensation, Social Security, unemployment, and disability insurance."[5]

Though stylized, our framework delivers a rich set of predictions that offer insight into the debate about the regulation of platform work. For instance, our comparative statics highlight the importance of beliefs about workers' valuations of benefits and how such beliefs interact with externalities, equity concerns and bargaining power in shaping the optimal policy.

By generating predictions about how optimal regulatory policies change as fundamentals in the economy shift, we complement recent work on the declining labor share and worker bargaining power discussed in Autor *et al.* (2020a) and Summers and Stansbury (2020), respectively. Specifically, our results highlight how a regulator might optimally adjust the contracting environment in order to maximize welfare in light of these recent macro trends. This can help explain recent pushes for increased minimum wages and improved working standards. It can also be used to highlight the value of soft power that the U.S. Congress applies to large firms, which can be seen as a means of shifting the default in terms of either quality or wages (Stewart and Stanford, 2017).

Related literature. This paper contributes to the literatures on contracting and implementation with renegotiation, and on delegated decision-making in organizations.

Our model is cast in the incomplete contracts framework of Hart and Moore (1988). The two parties in our model have observable information that is unverifiable by a third party (in our case, the regulator). Several papers have studied the role of renegotiation in incomplete contracts (Hermalin and Katz, 1991; Green and Laffont, 1992; Rubinstein and Wolinsky, 1992). Closest in spirit to our paper is Aghion *et al.* (1994), which focuses on how an initial contract might include provisions that govern the ex-post renegotiation process.[6]

[5]Assembly Bill 5 was signed into law in September 2019 only to be largely overridden by a statewide referendum, Proposition 22, in November 2020.

[6]A paper that makes a similar point to Aghion *et al.* (1994) is Chung (1991).

“Renegotiation design," as the authors describe it, has two elements: it specifies a “default outcome" in the event of a disagreement, and it specifies an allocation of bargaining power to one or the other contracting party. By contrast, in the standard incomplete contracts framework without renegotiation design, “no trade" is the only possible “default outcome." In the language of the US legal system, standard incomplete contracts frameworks focus on “at-will contracts" while renegotiation design allows for “specific performance contracts."

Our analysis extends Aghion *et al.*'s and differs in interpretation and emphasis. Rather than supposing that the two parties to a contract write their ex-post renegotiation provisions (default outcomes, allocation of bargaining power) into the initial contract, we assume that there is a regulator or social planner who has the authority to set and enforce default outcomes (and we treat bargaining power as exogenous). This shift in interpretation reflects that we are studying not the provisions of bilateral contracts themselves, but rather how lawmakers regulate bilateral contracts in the interests of the public. Our analysis gives us a rich description of how the optimal defaults change with different planner preferences over efficiency, equity and externalities, whereas Aghion *et al.* focus on whether, with some optimally-specified default and allocation of bargaining power, efficient investment is possible.

In providing foundations for the default mechanism, we also contribute to the literature on Nash implementation (Maskin, 1999; Moore and Repullo, 1988). In this context, the possibility of renegotiation restricts the set of implementable outcomes—any outcome that is not on the agents' Pareto frontier will be renegotiated to realize a Pareto improvement. Maskin and Moore (1999) fully characterize implementability when renegotiation cannot be prevented. Extending their analysis, we add a detailed characterization of implementable outcomes when the planner has preferences over the distribution of surplus between the agents.

The model can also be viewed through the lens of delegation. In fact a special case of our model reduces to a standard delegation problem as introduced by Holmström (1977, 1984) and generalized in Alonso and Matouschek (2008). When the regulator cares only about

externalities, then only the “quality" dimension of the contract is relevant, and the worker and firm can be thought of as a single agent maximizing the two parties’ joint surplus.

Much of the existing delegation literature focuses on delegation to a single agent. Our model analyzes delegation to *two* agents. While Martimort and Semenov (2008) considers a two-agent delegation problem in a legislative context when agents are asymmetrically informed, we assume agents have symmetric information. Delegation differs from standard mechanism design problems in that it assumes neither the principal nor any third parties (as in Tirole (1986) and Laffont and Martimort (1998)) can collect or disburse transfers to or from the agents. In the analysis in section 3.4, we assume the principal can commit to decision rules, and therefore this analysis differs from related cheap talk models like Krishna and Morgan (2001).

Our model takes a particular perspective on what it means to delegate a decision to two agents. We assume that the agents bargain efficiently over outcomes in the delegation set and that the principal can anticipate the outcome of agents’ bargaining. Although our analysis in section 3.4 begins in a territory covered by the Revelation Principle, it departs from the premises of the Revelation Principle when it introduces costly communication and the requirement of message-independence. Our analysis thus contributes to the study of tradeoffs between decentralization and centralization in organizational economics and mechanism design, thoroughly surveyed in Mookherjee (2006).

A common motivating example in the delegation literature is the regulation of a monopolist à la Baron and Myerson (1982). In the extant delegation literature, the results about the optimality of interval delegation are used to make sense of the widespread use of price caps in regulatory settings. Our model offers a new interpretation of regulation-as-delegation: when the government’s social welfare function incorporates distributional concerns, optimal regulation takes into account the bargaining process between a firm and its stakeholders.

3.2 Examples: Incomplete Contracts and Social Welfare

There is a central dichotomy in contract law between “default" rules and “immutable" rules. Default rules affect the contracting outcome without constraining the set of potential contracts. Immutable rules prevent the enforcement of specific contracts. We analyze the rationale behind these distinct tools and highlight how such rationales may be applied to regulation more broadly. We discuss three motivating examples. The first two examples—commercial contracts and marriage contracts—build toward our main example: the regulation of platform work.

Example 1 (Commerce) *The Uniform Commercial Code and “Reasonable” Defaults.*

We first discuss default delegation in the context of the Uniform Commercial Code (UCC). Here, default delegation is a response to a canonical incomplete contracting problem. The UCC’s primary objectives are to “simplify, clarify and modernize the law governing commercial transactions; to permit continued expansion of commercial practices through custom, usage and agreement of the parties; to make uniform the law among the various jurisdictions." In other words, this law is intended to ease and standardize contracting environments between commercial actors.

One of the key features of this standardization is to make clear how courts will enforce contracts. There are two aspects of enforcement: 1) a statement of contract terms which courts *will not* enforce (immutable rules) and 2) a statement of terms the court will enforce if the parties do not specify otherwise (default rules).

Consider a simple example where a buyer and a seller enter into a delivery contract for a widget. The contract specifies the time to delivery q as well as the price c. The seller invests a fixed cost of k to produce the good. There is some uncertainty about an ex-ante indescribable state of the world (ω, θ) which influences the buyer and sellers’ valuations of the contract. The state of the world has a component $\omega \in \{\omega_1, \omega_2\}$ which is ex-post verifiable (even though it is ex-ante indescribable). The other component of the state $\theta \in \{\theta_1, \theta_2\}$ is unverifiable. All together, the state could represent a variety of exogenous

factors that influence delivery times, some of which are verifiable (e.g. a container ship blocking the Suez Canal).

As in Aghion *et al.* (1994) the buyer and seller could agree to an initial contract (q_0, c_0) from which they can negotiate once ω and θ are revealed. Suppose (ω_1, θ_1) is a "bad" state of the world in which it is prohibitively costly for the seller to deliver at the pre-specified time q_0. When ω_1 occurs, the buyer and seller can renegotiate the contract to some $h((q_0, c_0), k, \omega_1, \theta_1)$, where h is an arbitrary bargaining function. But, there is a canonical hold up problem: there may be no guarantee that the seller can recoup its fixed cost k in the bad state (ω_1, θ_1). This sequence is shown in Figure 3.1a.

The Uniform Commercial Code (UCC) default rules governing delivery times offer the buyer and the seller a different option. The UCC states that "time for shipment or delivery... if not provided in this article or agreed upon shall be a reasonable time." That is, if the buyer and seller choose not to specify a delivery time in their initial contract, the UCC fills in the gap with a "reasonable" delivery time default rule. Crucially, when the gap is filled in, the regulator *will use all verifiable information* that is available at that time, including the realization of ω to determine what is reasonable: the default rule is a function of the verifiable portion of the state $(q(\omega), c(\omega))$. When the parties renegotiate in the state (ω_1, θ_1), they arrive at $h((q(\omega_1), c(\omega_1)), k, \omega_1, \theta_1)$. The reasonable default $(q(\omega_1), c(\omega_1))$ can be chosen to guarantee that the seller recoups its fixed cost k, even in the "bad" state. This case is shown in Figure 3.1b.

To summarize, the default rules can overcome the hold up problem in incomplete contracting problems when there is an ex-ante indescribable but ex-post verifiable component of the state ω. The government has the power to commit to enforcing a "reasonable" contract after the true state of the world is revealed, to the extent that the state is verifiable. This commitment allows the contracting parties to avoid the disaster case for the seller (ω_1, θ_1), which might otherwise jeopardize the contract. Furthermore, as shown in Appendix C.1, the regulator can achieve this by ensuring a distribution of surplus consistent with the ex ante bargaining position of the agents. In this sense the regulator provides a default that the

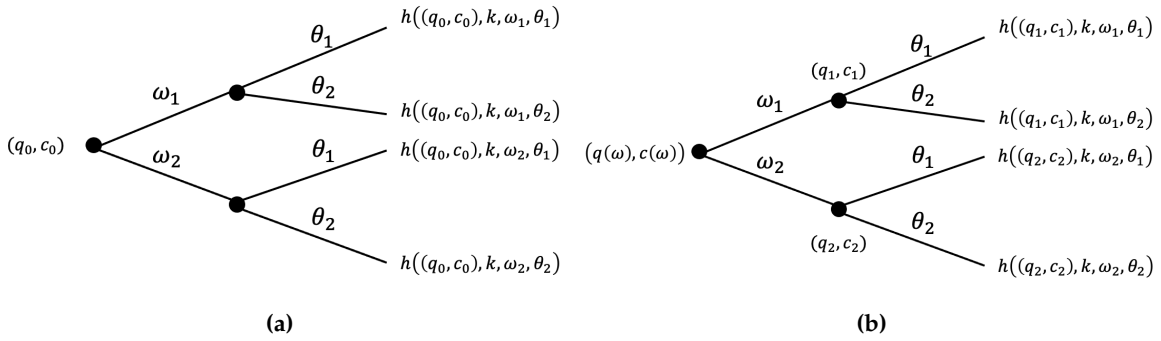


Figure 3.1: *(a) Bargaining from an Explicit Prespecified Contract; (b) Bargaining from the U.C.C.'s "Reasonable" Default*

agents "would have wanted" by ensuring a particular distribution of surplus. Our analysis in section 3.4 will generalize this example.

Example 2 (Marriage) *The "Equitable Division" Default and Limited Enforcement of Premarital Contracts.*

Another setting which has fallen under the purview of a uniform act in the U.S. is the law governing premarital agreements. The Uniform Premarital Agreement Act, adopted (often with slight differences) by 28 states, specifies the default marriage contract which holds in the absence of an alternative premarital agreement. Beyond the default rules outlined in the UPAA, states can also decide the extent to which they will enforce particular alternative contracts.

In the case of marriage, the government explicitly considers issue of fairness when choosing default and immutable rules (Bix, 1998). When it comes to the division of assets, divorce laws in the majority of states enforce an "equitable division" default (Hersch and Shinall, 2019). For example, in Mass. Gen. Laws ch. 208 § 34 (2018), the division of assets upon divorce is decided through a holistic evaluation of "the length of the marriage, the conduct of the parties during the marriage, the age, health, station, occupation, amount and sources of income, vocational skills, employability, estate, liabilities and needs of each of the parties."

However, the division of assets in the event of divorce is often written into enforceable

pre-marital agreements. That is, spouses can contract around “equitable division" default rules by explicitly specifying how assets are to be divided in the event of divorce. The Uniform Premarital Agreement Act puts few limitations on pre-martial contracts, with the primary exceptions being: a) the agreed division of assets cannot be “extremely unfair" and involve a lack of disclosure, and b) the agreed division of assets cannot necessitate government assistance for one of the spouses. Courts will also not enforce any premarital contract term related to the division of custody. These limits on premarital contracts are immutable laws that specifically address the presence of externalities which affect the government and potential children.

Recall that in the previous example, the U.C.C. supplies a “reasonable" default delivery time when not otherwise specified, at least in part to ensure particular distributions of surplus in ex-ante noncontractible states of the world. In marriage contracts, the “equitable division" default rules serve a similar purpose, however, in this case, the government has explicit distributional concerns for fairness reasons.

It is telling to put these default rules in historical perspective. Until the mid-1970s, most states refused to enforce *any* pre-marital contracts. There was a widely held view, at the time, that premarital agreements were unfair because they were designed to “protect the wealth and earnings of an economically superior spouse from being shared with an economically inferior spouse" (Bix, 1998). As social facts and attitudes have shifted, the government has increased the scope for bargaining while maintaining a default which is predicated on equity. Meanwhile, in instances where there are outside parties involved (the government or children), the government has maintained some immutable terms. Thus, the regulation of marriage illustrates how notions of equity and externality naturally lead to the application of defaults and restricted delegation sets, respectively.

Example 3 (Work) *The Classification of Platform Workers.*

The last two examples lean on the presence of (or potential for) incomplete contracts as a reason for government intervention in the form of default and immutable rules. This paper argues that this method of regulating contracts is more general. The regulation of bilateral

contracts more generally can be viewed through the lens of default delegation: default rules affect the disagreement outcome of agents' bargaining in order to affect the distribution of surplus across unverifiable states of the world, while limits on enforceable contracts curb externalities.

To show the generality of the logic of default delegation, we will focus on the example of regulating app-based platform workers, which has led to contentious debate in the U.S. and abroad. The debate has largely focused on whether these workers should be classified as employees or independent contractors. As independent contractors, there are very few limitations on the contracting space. On the other-hand, employee status entails a minimum suite of benefits as well as firm contributions for programs such as social security, unemployment insurance, payroll taxes, and premiums for worker's compensation.

Thus, the imposition of employee status, constrains the set of contracts that the government enforces. More subtly, the change in employment status also affects the workers' outside options, and thus their disagreement outcome in employment negotiations. Employee status raises the minimum levels of benefits and wages that workers' can expect to receive from a different company. In this sense the government can indirectly influence a worker's outside option.

One way of understanding what has happened in the rise of platform work is that some ex-ante indescribable state of the world ω was realized, where ω was a unique work arrangement that could not have been foreseen. In the absence of reclassificaiton, the workers in this work arrangement remained classified as independent contractors with the default defined by (q_i, c_i). Despite being distinct from traditional independent contractors, these workers are only able to negotiate a contract based on independent contractor status resulting in $h(q_i, c_i, \tilde{\omega})$. If the firm has all of the bargaining power, any negotiation will lead to a distribution of surplus that favors the firm over the workers. This can be especially detrimental to workers in situations where there may have been up-front investments in gig-work such as quitting full-time employment, purchasing a vehicle, etc. In this case, the worker could be viewed as suffering from a "hold up" problem. While this version of the

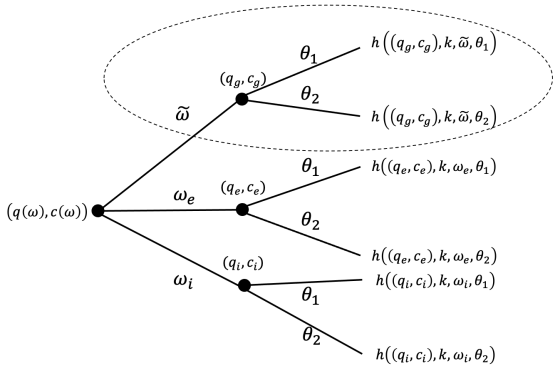


Figure 3.2: *An Unforeseen Work Regime Requiring a New Default*

"hold-up" problem may not be so detrimental as to prevent workers from entering in the first place, it may prevent the government from achieving a desired distribution of the surplus from the agreement.

The goal of the regulation of employment contracts is to update worker classification in light of these changing circumstances. In the U.S., the two main classifications of work arrangements are independent contractors and employees, which can be understood as coming along with contracts (q_i, c_i) and (q_e, c_e). Labor regulations are thus analogous to default and immutable rules. They do not fill gaps in incomplete contracts, but they do use verifiable information available at a given time to smooth the distribution of surplus across states, as well as protect parties within and outside the contract.

Any attempt to reclassify workers can thus be understood as a drive to change the default contract in light of an ex-ante indescribable but ex-post verifiable state of the world. Ex-ante, workers, employees, and the government had not foreseen the unique working arrangements $\tilde{\omega}$ that would arise out of the spread of platform work. So, what regulations of platform work seek to do is to update the default contract to some $(q_d(\tilde{\omega}), c_d(\tilde{\omega}))$, where the default is contingent on verifiable information.

The search for a better classification of platform work can thus be understood as the search for a more appropriate default contingent on new information. Furthermore, the appropriate default depends on the government's preferences over efficiency, equity and

externalities. To see this, recall California's 2019 Assembly Bill No. 5 (AB5). The goal of the bill was "to ensure workers who are currently exploited by being misclassified as independent contractors instead of recognized as employees have the basic rights and protections they deserve under the law, including a minimum wage, workers' compensation if they are injured on the job, unemployment insurance, paid sick leave, and paid family leave." In the language of our model, AB5 argued that $q_d(\tilde{\omega}), c_d(\tilde{\omega})$ should be changed from (q_i, c_i) to (q_e, c_e).

After the passage of AB5, Lyft and Uber spent more than $200MM in order to pass a referendum, Proposition 22, which returned the employment classification of app-based drivers to independent contractors. The proposition made the argument that the classification of employee was not desirable for workers who placed a high value on flexibility. Proposition 22 did however include some minimum earnings requirements, non-discrimination protections and some minimum health insurance.[7] In some sense, Proposition 22 thus resulted in a new form of default contract, (q_g, c_g) which is a new classification tailored to the state of the world $\tilde{\omega}$. See Figure 3.2 for an illustration of these contracts.

The rhetoric surrounding the debates over Proposition 22 were largely focused on questions of equity versus efficiency.[8] Those that opposed Proposition 22 argued that the classification of workers as employees would improve their expected outcomes by entitling them to benefits and wage standards. Classifying workers as employees would lead to a more equitable split of the total surplus generated from these contracts. The proponents of Proposition 22 argued that platform workers have a relatively low valuation for employment benefits that come with employee status and profit from being able to accept contracts with lower benefits and higher pay. While there are aspects of the "state of the world" that are unverifiable—such as the degree to which workers in different companies *actually* value

[7] https://pmcinsurance.com/blog/how-assembly-bill-5-differs-from-proposition-22/

[8] See, for example, https://www.americanactionforum.org/insight/whats-next-for-prop-22-and-debates-around-independent-workers/

benefits relative to cash—there are aspects of the state that are verifiable that the government can optimally condition its worker classification upon.

3.3 Model

To fix ideas, we present our general model in terms of an employment contract between a worker and a firm. The firm and the worker, more generally, can be understood to be any two parties that are negotiating the terms of a bilateral contract. In line with the delegation perspective of our model, we sometimes refer to the firm and the worker collectively as the "agents," while the regulator in this setting is the "principal." The principal aims to maximize social welfare by influencing the contracting environment and choosing which contracts to enforce. The state of the world, which governs agent preferences (and thus social welfare), is observable but unverifiable—so the principal is limited in that it cannot enforce contracts that are contingent on the state. We also rule out the possibility that the principal can make monetary transfers to or from the agents, as is standard in the delegation literature.

Contracts. Agents bargain over the terms of a contract $(q, c) \in \mathcal{Q} \times \mathcal{C} \subseteq \mathbb{R}^2$. The term q in the contract represents a dimension over which agents have possibly state-dependent valuations. We often refer to q as "quality." For example, q might represent a particular benefit in an employment contract, such as the degree of health insurance coverage provided by the firm to the worker. Meanwhile, c captures the money that will be transferred from one agent to the other. In an employment contract, c is the compensation (salary, wages) paid to the worker by the firm.

Preferences. The regulator and the agents' utilities depend on the negotiated contract and on the state of the world. The state of the world is $\theta \in \Theta \subset \mathbb{R}$.

The two agents, the firm (f) and the worker (w), both have quasilinear state-dependent utility functions over the outcome (q, c):

$$\textbf{Firm: } U_f(q, c; \theta) = u_f(q; \theta) - c \quad \textbf{Worker: } U_w(q, c; \theta) = u_w(q; \theta) + c.$$

The principal's goal is to maximize a generalized social welfare function,

$$\text{SWF}(q, c; \theta) = \text{SWF}(U_f, U_w, U_r; \theta) = \underbrace{U_f + U_w}_{\text{"efficiency"}} - \underbrace{\beta(U_f - U_w)^2}_{\text{"equity"}} + \underbrace{\gamma U_r(q; \theta)}_{\text{"externality"}}$$

where $\beta \geq 0$ and $\gamma \geq 0$ scale the magnitude of the social cost associated with equity and externalities, respectively. Equity concerns are represented by a quadratic penalty. The equity term is maximized when the worker and firm attain the same utility. This "equal split" equity objective is assumed to simplify exposition.[9] Externalities are represented by $U_r(q; \theta)$. For example, in an employment contract, the government may end up paying for health care that is not covered in an employer provided health insurance program. This may be more or less costly depending on the risks of the occupation.

The generalized social welfare function nests more specific social welfare functions. For instance, when $\beta = 0$ and $\gamma = 0$, the regulator's objective simplifies to maximizing efficiency. When $\beta = 0$ and $\gamma > 0$, the regulator additionally attempts to internalize externalities that may arise out of the contract. When $\beta > 0$ and $\gamma = 0$, the regulator trades-off efficiency losses with equity gains.

Information. The agents have complete and symmetric information about the state θ. However, the state is not verifiable by the principal. The principal's beliefs over the state are represented by the cumulative distribution function $G(\theta)$ with probability density function $g(\theta)$. All other details of the environment are common knowledge among the principal and the agents, including the outcome of bargaining.

Bargaining. Agents Nash bargain over the terms of the contract (q, c) given an outside option $d = (q_d, c_d)$ and the state of the world θ. The worker has Nash bargaining weight δ while the firm has weight $(1 - \delta)$. These weights, in the case of employment negotiations, are influenced by the presence or absence of a union, collective bargaining protections, as

[9] It is straightforward to verify that the results presented here hold for broad set of equity objectives, i.e. any particular desired distribution of surplus x and $1 - x$. A discussion of the form of the inequity penalty appears in subsection C.4.1.

well as employer concentration and monopsony power. The default contract which obtains in the event of a disagreement is d. That is, the bargained contract is given by

$$h(d;\theta) \equiv \arg\max_{q\in\mathcal{Q},c\in\mathcal{C}} (U_w(q,c;\theta) - U_w(d;\theta))^{\delta}(U_f(q,c;\theta) - U_f(d;\theta))^{1-\delta}. \tag{3.1}$$

We write $\tilde{h}$ to refer to the maximand in (3.1). Here $\mathcal{Q}$ and $\mathcal{C}$ refer to the set of enforceable quality levels and transfers, respectively.

The regulator's problem. The regulator's goal is to maximize social welfare. Section 3.4 will consider a variety of different assumptions about the degree to which the regulator is constrained in its choice of social welfare-maximizing mechanisms.

Default delegation. The analysis in Section 3.4 leads to a particular indirect mechanism that we call *default delegation*. In default delegation, the regulator uses two tools to affect the contracting environment: (1) it chooses the *delegation set* $\mathcal{Q}$ of contracts it is willing to enforce and (2) it sets the *default outcome* outcome d which serves as a disagreement point in the agents' negotiations. That is, the regulator maximizes

$$\max_{\mathcal{Q},d} \mathbb{E}_{\theta}[\text{SWF}(q^*(\theta), c^*(\theta);\theta)] \tag{3.2}$$

with

$$(q^*(\theta), c^*(\theta)) \equiv \arg\max_{q\in\mathcal{Q},c\in\mathcal{C}} \tilde{h}(d,\theta).$$

In section 3.4, we will discuss the properties of this indirect mechanism in depth, comparing its performance to an unconstrained direct mechanism. In section 3.5, we assume that the regulator uses a default delegation mechanism, and analyze the optimal policy choice.

Timing of default delegation. The regulator chooses a delegation set and default outcome $\{\mathcal{Q}, d\}$. Then with common knowledge of the true state θ, the agents bargain over the terms of the contract (q, c), with d serving as the disagreement outcome, and $\mathcal{Q} \times \mathcal{C}$ defining the feasible bargaining outcomes. Then the contracts are signed and utilities are realized.

3.4 First Best Analysis

In this section, we characterize the set of contracts that are implementable via default delegation. We begin by noting that first-best contracts are trivially implementable with default delegation when the regulator is concerned only about the efficiency of the contract. Then, we build off the approach in Maskin and Moore (1999) in order to show when it is without loss for the regulator to choose default delegation in the presence of equity and efficiency concerns. The main departure from Maskin and Moore is that we derive explicit conditions on implementability assuming that agents engage in Nash bargaining, and that the regulator has specific preferences over the distribution of surplus and externalities captured by $\beta > 0$ and $\gamma > 0$, respectively. Adding this structure allows us to investigate the implications of Maskin and Moore's general implementation theorems for specific applications of interest.

In other words, the goal of this section is to understand when the regulator can achieve the first-best outcome with default delegation for different social welfare functions. We define first-best as follows.

Definition 4 (First Best Outcomes) *The first best outcomes are* $(q^*, c^*) = \{(q_\theta, c_\theta)\}_{\theta \in \Theta}$ *where*

$$(q_\theta, c_\theta) \in \arg\max_{(q,c)} \text{SWF}(q, c; \theta)$$

for $\theta \in \Theta$.

3.4.1 Implementation of efficient contracts

We consider different social welfare functions in turn, beginning with the most familiar one based only on (ex-post) efficiency. When $\beta = 0$ and $\gamma = 0$, there are neither equity nor externality concerns, and the regulator's goal is only to implement the efficient contract. In particular, the regulator does not have a preferred distribution of surplus, and thus each state corresponds to a *set* of first-best contracts (q_θ, c_θ) where c_θ is any value in $\mathcal{C}$ and q_θ is the efficient level of quality.

Since bilateral bargaining is efficient, and the regulator is indifferent about the transfers c that result from the bargaining process, the first-best outcome is achieved with default delegation as long as the regulator is willing to enforce the first-best quality level in each state (i.e. $q_\theta \in \mathcal{Q}$).

Lemma 1 *Assume $\beta = 0, \gamma = 0$. Then if $q_\theta \in \mathcal{Q}$ for all θ, the first-best is implementable with default delegation for any choice of d including the status quo.*

In this case, there is no reason for the regulator to constrain the set of enforced outcomes $\mathcal{Q}$ nor to judiciously select a default outcome d. With no intervention, the parties reach the first-best outcome.

3.4.2 Default delegation with equity concerns

It may be that in order to achieve the first best level of quality q_θ, the transfers that result from agents' bargaining c_θ constitute an undesirable distribution of the surplus. In our examples, the undesirability of a particular distribution could be derived from ex ante contracting efficiency (as in incomplete contracts, see Example 1 in section 3.2) or from fairness or other social concerns about distribution (as in marriage and employment contracts, see Examples 2 and 3 in section 3.2). Either way, when the regulator has equity concerns, the first best in state θ is a single pair (q_θ, c_θ) where q_θ is the efficient quality level and c_θ evenly distributes the surplus between the two parties (this particular first-best distribution is due to our simplified social welfare function, but as noted in subsection C.4.1, the results hold for general objectives regarding the distribution of surplus).

In this subsection we begin by characterizing the full set of implementable contracts when the regulator has equity concerns and can design *any* mechanism. We show that under some circumstances, any outcome implementable with a general direct mechanism can also be achieved through default delegation. That is, the regulator can do just as well with a decentralized default mechanism as it can with an arbitrary centralized direct mechanism.

We begin by characterizing the full set of implementable outcomes when the regulator cannot prevent the parties from renegotiating, using the results in Maskin and Moore (1999).

The regulator's problem can be translated into an optimization over incentive compatible direct mechanisms in which both parties submit reports of the state θ.

The regulator's problem: Direct mechanism. In the direct mechanism, the action space is Θ^2. We refer to the worker's report as $\hat{\theta}_w$ and the firm's report as $\hat{\theta}_f$, resulting in a profile of agents' reports $\hat{\theta} = (\hat{\theta}_w, \hat{\theta}_f)$. We will begin by assuming that the regulator will enforce any contract, $\mathcal{Q} = \mathbb{R}$. The regulator's problem is to maximize

$$\max_{g(\hat{\theta})} \mathbb{E}_\theta[\text{SWF}(h(g(\hat{\theta}), \theta); \theta)] \tag{3.3}$$

subject to incentive compatibility constraints

$$U_i(h(g(\theta, \theta), \theta) \geq U_i(h(g(\theta', \theta), \theta)$$

for all $\theta, \theta' \in \Theta$. The following proposition characterizes first-best implementable outcomes.

Proposition 1 (Maskin and Moore 1999) *Assume that the Pareto frontier is linear in all states $\theta \in \Theta$. The first-best is implementable in Nash equilibrium and any refinement if and only if there exists a function $g : \Theta \times \Theta \to \mathcal{Q} \times \mathcal{C}$ such that*

(i)

$$h(g(\theta, \theta), \theta) \in \arg\max_{g(\theta,\theta)} \text{SWF}(h(g(\theta, \theta), \theta),$$

(ii) and for $i \in \{w, f\}$

$$U_i(h(g(\theta, \theta), \theta)) \geq U_i(h(g(\theta', \theta), \theta)) \tag{3.4}$$

for all $\theta, \theta' \in \Theta$.

In general, default delegation will be restrictive in the sense that the regulator could implement more outcomes if any direct mechanism were available. However, when there are only two states, default delegation can replicate any implementable outcome under mild conditions on agents' utility functions. We proceed by discussing the two state case in depth.

Direct mechanism: Two states. To make Proposition 1 more concrete, consider the two state case where $\Theta = \{\theta_l, \theta_h\}$. Here, we write the first best outcomes to be (q_l, c_l) and (q_h, c_h) in state θ_l and θ_h, respectively. Proposition 1 says that the first best is implementable in any refinement of Nash equilibrium as long as neither player would deviate in the reduced form game shown in Table 3.1.

Table 3.1: *Direct mechanism for implementing (q_θ, c_θ) in state θ.*

	$\hat{\theta}_f = \theta_l$	$\hat{\theta}_f = \theta_h$
$\hat{\theta}_w = \theta_l$	(q_l, c_l)	(q_{lh}, c_{lh})
$\hat{\theta}_w = \theta_h$	(q_{hl}, c_{hl})	(q_h, c_h)

The regulator's problem then is to choose a mechanism $g(\hat{\theta})$ so that the only equilibria of the game are (q_θ, c_θ) for $\theta \in \Theta$. This amounts to a selection of off-path "threats" (q_{hl}, c_{hl}) and (q_{lh}, c_{lh}).

In characterizing implementable outcomes, it will be useful to introduce notation for a difference operator $\Delta(u, q, \theta) \equiv u(q_\theta, \theta) - u(q, \theta)$. The off-path threats must satisfy

$$(1-\delta)[\Delta(u_w, q_{mn}, \theta_n) - \Delta(u_w, q_{mn}, \theta_m)] - \delta[\Delta(u_f, q_{mn}, \theta_n) - \Delta(u_f, q_{mn}, \theta_m)] \geq c_m - c_n \quad (3.5)$$

for $m, n \in \{l, h\}$ with $m \neq n$. This condition is very closely related to the cross partials of the agents' utility with respect to q and θ, which will be useful in building intuition below.

Default delegation: Two states. How does default delegation differ from the general direct mechanism? Default delegation is decentralized, in the sense that it is message-independent. It does not require a solicitation of agents' reports.

To understand what is possible with message-independence, let's first consider a restriction to games in which only the off-path threats are message-independent in the sense that there is only one outcome for when the agents' reports disagree. That is, $g(\theta, \theta') = d$ for $\theta \neq \theta'$. This game is represented in Table 3.2.

Table 3.2: *Direct mechanism with message-independent threats* ($|\Theta| = 2$).

	$\hat{\theta}_f = \theta_l$	$\hat{\theta}_f = \theta_h$
$\hat{\theta}_w = \theta_l$	(q_θ, c_θ)	(q_d, c_d)
$\hat{\theta}_w = \theta_h$	(q_d, c_d)	(q_θ, c_θ)

Here, (q_d, c_d) must satisfy all the constraints characterized by (3.5). This implies that these constraints must hold with equality. Since they hold with equality, for well-behaved U_i it must be that $h((q_d, c_d), \theta) = h((q_\theta, c_\theta), \theta) = (q_\theta, c_\theta)$.

Therefore, the game that implements first-best with message-independent threats in Table 3.2 results in the same outcome as default delegation. The next proposition characterizes when it is without loss for the regulator to choose the fully decentralized, message-independent default delegation mechanism.

Proposition 2 *Assume that* $|\Theta| = 2$. *If the first-best is implementable and* U_f *and* U_w *are continuous, then the first-best is implementable with default delegation. The default* $d = (q_d, c_d)$ *satisfies*

$$(1-\delta)[\Delta\Delta(u_w, q_d, \theta_h, \theta_l)] - \delta[\Delta\Delta(u_f, q_d, \theta_h, \theta_l)] = c_l - c_h \tag{3.6}$$

where $\Delta\Delta(u, q, \theta, \theta') \equiv \Delta(u, q, \theta) - \Delta(u, q, \theta')$.

As mentioned above, the cross-partials of the worker and firm utility functions with respect to q and θ give intuition for the logic behind the condition in Proposition 2. The following corollary contains a sufficient condition for implementation with default delegation in the absence of externality concerns:

Corollary 1 *Assume that* $|\Theta| = 2$ *and* $\gamma = 0$. *If there exists* $x \in \mathbb{R}_+$ *such that at all points* (q, θ)

$$\left|(1-\delta)\frac{\partial^2 U_w}{\partial q \partial \theta} - \delta\frac{\partial^2 U_f}{\partial q \partial \theta}\right| > x \tag{3.7}$$

then the first-best is implementable via default delegation.

This implies that if the worker has supermodular preferences and the firm has submodular preferences then there exists a contract (q_d, c_d), which implements the first best. Furthermore, these cross-partials provide insight into the feasibility of specific transfers with different bargaining parameters. Specifically, corollary 2 shows conditions under which the government may find it less difficult to achieve a desired distribution of surplus:

Corollary 2 *If the worker and firm have supermodular and submodular preferences, respectively such that*

$$\frac{\partial^2 U_w}{\partial q \partial \theta} = b > 0 \quad \textit{and} \quad \frac{\partial^2 U_f}{\partial q \partial \theta} = -a < 0$$

with $b > a$, the regulator is able to achieve the same difference in transfer, $c_l - c_h$ with a lower default quality level, q_d, when the workers have less bargaining power.

In other words, corollary 2 states that when the workers' preferences over q are more responsive to the state θ, the regulator can achieve a broader set of transfers $\{c_l, c_h\}$ when the firm has higher bargaining power. More generally, the regulator has more scope for achieving particular distributions of surplus when the party with more state-dependent preferences has less bargaining power.

As discussed in Maskin and Moore (1999), when the regulator enforces any contract written between agents $\mathcal{Q} = \mathbb{R}_+$, this significantly constrains the set of implementable outcomes. Furthermore, we could imagine that the regulator can choose to only enforce contracts where $(q, c) \in \{(q_l, c_l), (q_h, c_h), d\}$ noting that the default d must be enforceable for the negotiation to be well defined but can itself equal (q_l, c_l) or (q_h, c_h). In this case, the planner's problem becomes:

$$\max_{\{(q_l,c_l),(q_h,c_h),d\}} \mathbb{E}_\theta[\text{SWF}(q^*(\theta), c^*(\theta); \theta)] \tag{3.8}$$

with

$$(q^*(\theta), c^*(\theta)) \equiv \underset{(q,c)\in\{(q_l,c_l),(q_h,c_h),d\}}{\arg\max} (U_w(q, c; \theta) - U_w(d; \theta))^\delta (U_f(q, c; \theta) - U_f(d; \theta))^{1-\delta}.$$

Allowing the regulator to constrain the set of enforceable contracts, we get the following

result:

Proposition 3 *The first-best is implementable with constrained default delegation if there exists a default* $d = (q_d, c_d)$ *such that*

$$(U_w(q_\theta, c_\theta; \theta) - U_w(d; \theta))^\delta (U_f(q_\theta, c_\theta; \theta) - U_f(d; \theta))^{1-\delta} > \tag{3.9}$$
$$(U_w(q_{\theta'}, c_{\theta'}; \theta) - U_w(d; \theta))^\delta (U_f(q_{\theta'}, c_{\theta'}; \theta) - U_f(d; \theta))^{1-\delta}$$

for all $\theta, \theta' \in \Theta$.

Before the regulator needed to ensure agents' incentive compatibility conditional on the transfer based on their optimal division of surplus. Now, the regulator can refuse to enforce contracts which they would optimally choose to bargain to. This generally makes the set of implementable contracts broader. However, an implication of corollary 1 is that if (3.7) is satisfied then the regulator does not expand the set of implementable outcomes by choosing to limit the set of enforceable contracts. In other words, when the regulator has distributional concerns, it does not increase welfare by restricting the set of enforced contracts when (3.7) holds. We summarize this observation in the following Corollary.

Corollary 3 *Assume that* $|\Theta| = 2$ *and* $\gamma = 0$. *If the condition from corollary 1 is satisfied, then the regulator cannot increase the set of implementable outcomes by limiting the set of enforceable contracts.*

3.4.3 Default delegation with externality concerns

So far, this section has considered cases in which the regulator has preferences over efficiency and equity alone. In such cases, it is never beneficial for the regulator to restrict the set of enforced outcomes $\mathcal{Q}$ because the agents' bargaining always achieves the efficient q_θ. When the regulator has only equity and efficiency concerns, it influences the distribution of surplus by choosing the default $d = (q_d, c_d)$ which serves as a disagreement outcome in agents' bargaining.

Now, we include social welfare functions where externalities affect the first-best outcomes. In such cases, the first best (q_θ, c_θ) will not be on the agents' Pareto frontier, and so the regulator will not be able to attain first best without restricting the set of enforced contracts.[10]

If the externality depends on the quality level such that $\frac{\partial U_r}{\partial q} \neq 0$ and $\gamma > 0$, the regulator may still be able to achieve first best by delegating the choice to the agents, and letting them choose from a reduced set of contracts that contain the optimal quality levels corresponding to the first-best outcomes in each state. In other words, the regulator can achieve first best via default delegation by setting the delegation set to be a discrete set $\mathcal{Q} = \{q_\theta\}_{\theta \in \Theta}$. This delegation set will achieve first best when the agents' total surplus in state θ is higher at the first best level of quality, q_θ, than at an alternative quality level $q_{\theta'}$.[11] The following proposition characterizes when first-best is implementable with default delegation when there are externality concerns $(\gamma > 0)$ but not equity concerns $(\beta = 0)$.

Proposition 4 *Assume that Θ is a discrete set and $\beta = 0$. Then the first-best is implementable via default delegation with $\mathcal{Q} = \{q_\theta\}_{\theta \in \Theta}$ if*

$$u_f(q_\theta, \theta) + u_w(q_\theta, \theta) \geq u_f(q_{\theta'}, \theta) + u_w(q_{\theta'}, \theta) \tag{3.10}$$

for all $\theta, \theta' \in \Theta$.

Condition (3.10) is restrictive. It is easier to satisfy when the externality term is small. Furthermore, note that we are only allowing the regulator to limit the delegation set through quality. As discussed above, we could have allowed the regulator to limit the set of enforceable contracts such that we repeat the result of Proposition 3. This expands the set of implementable contracts because it is able to implement quality levels which do not

[10]Note that once we allow the regulator to restrict the set of enforced contracts, the framework in Maskin and Moore (1999) no longer applies. In Maskin and Moore the social planner cannot place any restrictions on renegotiation. We now assume that the social planner can restrict the set of enforceable quality levels $q \in \mathcal{Q}$. This allows us to study what happens when the planner can partially, but not completely, prevent renegotiation. This restriction on "quality" q but not transfers c is valuable as a theoretical exploration of the degree to which renegotiation constrains implementation. Furthermore, this restriction aligns with the perspective in Glaeser and Shleifer (2001), which posits that regulations that place limits on prices are more difficult to enforce because compliance is less observable.

[11]Recall that we have assumed that when the Nash bargaining solution is not in the feasible set, the agents select the point in the feasible set that maximizes their joint surplus.

satisfy equation 3.10 by making the distribution of surplus less favorable in particular states relative to the gains from increasing total surplus. However, as discussed in Glaeser and Shleifer (2001), the regulator might be limited in its ability to enforce particular transfers across states.

The next proposition adds back equity concerns $\beta > 0$. Essentially combining Propositions 2 and 4 we get the general result:

Proposition 5 *Assume that* $|\Theta| = 2$. *Then the first-best is implementable via default delegation if*

(i) there exists $d = (q_d, c_d)$ *that satisfies* (3.6), *and*

(ii) $\mathcal{Q} = \{q_\theta\}_{\theta \in \Theta}$ *satisfies* (3.10) *for all* $\theta, \theta' \in \Theta$.

The important takeaway from this section is that we can provide a mechanism design rationale for the distinction between immutable and default contract rules. The default rules are used in order to obtain a particular distribution of the surplus in different states of the world while immutable rules internalize externalities. Notice that Proposition 5 requires that the state space Θ has cardinality two. This is because Proposition 2 does not hold when there are more than two states. Now that we have discussed the ways in which default delegation allows a regulator to do the best it could do, we next turn to a discussion of the limitations of default delegation.

3.4.4 Limitations of default delegation

We first generalize Proposition 5 to more than two states in Proposition 6 below. Then we discuss the implications.

Proposition 6 *Assume* $|\Theta| > 2$ *and* Θ *is a discrete set. Then the first best outcomes* $\{(q_\theta, c_\theta)\}_{\theta \in \Theta}$ *are implementable with default delegation mechanism* $(d, \mathcal{Q})$ *if*

(i) $d = (q_d, c_d)$ *satisfies*

$$(1 - \delta)[\Delta\Delta(u_w, q_d, \theta, \theta')] - \delta[\Delta\Delta(u_f, q_d, \theta, \theta')] = c_{\theta'} - c_\theta \tag{3.11}$$

for all θ, θ', *and*

(ii) $\mathcal{Q} = \{(q_\theta, c_\theta)\}_{\theta \in \Theta}$ *satisfies* (3.10).

An implication of the first hypothesis (i) in this proposition is that default delegation *does not* implement the full set of implementable outcomes when there are more than two states, even when U_f and U_w are continuous in q and θ. To give intuition for why this is the case, we consider a setting with three states, i.e. when the state space is $\Theta = \{\theta_l, \theta_m, \theta_h\}$.

As before, it is useful to study the direct mechanism in which the regulator is restricted to using message-independent threats. In this case the game takes the form presented in Table 3.3 (which is a simple extension of Table 3.2 to three states).

Table 3.3: *Direct mechanism with message-independent threats* ($|\Theta| = 3$).

	$\hat{\theta}_f = \theta_l$	$\hat{\theta}_f = \theta_m$	$\hat{\theta}_f = \theta_h$
$\hat{\theta}_w = \theta_l$	(q_l, c_l)	(q_d, c_d)	(q_d, c_d)
$\hat{\theta}_w = \theta_m$	(q_d, c_d)	(q_m, c_m)	(q_d, c_d)
$\hat{\theta}_w = \theta_h$	(q_d, c_d)	(q_d, c_d)	(q_h, c_h)

In general, it will not be possible to find a (q_d, c_d) that implements the first-best outcomes. This is due to the constraints imposed by equation (3.6), which relates the differences in utility at q_θ and q_d in each state to the difference in desired transfers $c_h - c_l$. When there is a third state, there is now another value c_m, which appears in two other conditions analogous to (3.6). In order to implement the first-best, the value c_m must satisfy

$$(1-\delta)[\Delta\Delta(u_w, q_d, \theta_h, \theta_m)] - \delta[\Delta\Delta(u_f, q_d, \theta_h, \theta_m)] = c_m - c_h \tag{3.12}$$

$$(1-\delta)[\Delta\Delta(u_w, q_d, \theta_m, \theta_l)] - \delta[\Delta\Delta(u_f, q_d, \theta_m, \theta_l)] = c_l - c_m \tag{3.13}$$

where again $\Delta\Delta(u, q, \theta, \theta') \equiv \Delta(u, q, \theta) - \Delta(u, q, \theta')$. In general, it will not be the case that c_m satisfies both of these equations. So, there are outcomes (q^*, c^*) that are implementable with

a general direct mechanism that are not implementable when the regulator is restricted to message-independent threats. Thus, Proposition 2 does not hold when there are three states. In fact, as the following proposition highlights, Proposition 2 does not hold when $|\Theta| > 2$.

Corollary 4 *Assume $|\Theta| > 2$ and Θ is a discrete set. Then there are outcomes that are implementable in a general direct mechanism that are not implementable with default delegation, regardless of whether u_f and u_w are continuous in q, θ.*

The logic is that default delegation allows for two degrees of freedom which establish the level and difference between transfers across two states. With those degrees of freedom occupied there is no adjustment available for a third (or fourth or fifth...) state. Outcomes which may be implementable in a message-dependent mechanism in general will not be implementable in a default delegation mechanism.

To summarize, it is without loss to restrict to default delegation when there are only two states—Proposition 2 says that the entire set of implementable outcomes is implementable with default delegation under mild conditions. However, when there are more than two states, default delegation will, in general, meaningfully constrain the regulator's ability to achieve first-best relative to any unconstrained message-dependent direct mechanism.

Nonetheless, there are some scenarios that may be of interest in which results derived in the $|\Theta| = 2$ case extend to the $|\Theta| > 2$ case. One such situation is when the regulator has a maxmin objective function, which is discussed in Appendix C.3.

Beyond analyzing implementability, it is important to consider the important tradeoffs which occur when we extend this mechanism to many states. The regulator must consider efficiency, equity and externality both within a particular state and across states. With this in mind the next section looks at a particular example with a continuum of states in an attempt to gain intuition for how a regulator may adjust policies as underlying parameters evolve when using default delegation. But first we fix terms by interpreting the preceding analysis through the lens of our examples.

3.4.5 Interpreting results

The preceding analysis helps to clarify that the government aims to set or influence defaults in order to ensure particular distributions of surplus in unverifiable states. Meanwhile, the government sets limits on enforceable contracts in order to curb externalities. The key results are summarized in Table 3.4.

Table 3.4: *Summary of results: First-best implementation with default delegation*

Efficiency + Equity	Efficiency + Externalities	Efficiency + Equity + Externalities
(Proposition 2) achieved with d	(Proposition 4) achieved with $\mathcal{Q}$	(Proposition 5) achieved with $\{d, \mathcal{Q}\}$

We interpret the results of the first-best analysis in the context of the examples presented in section 3.2. Table 3.5 outlines examples of defaults and limits in commercial law, marriage law and labor law.

Table 3.5: *Examples of Defaults and Limits in Settings from Section 3.2*

#	Example	Default d	Limits $\mathcal{Q}$
1	**Commerce**	"reasonable" times	no "manifestly unreasonable" delivery times
2	**Marriage**	"equitable division" of assets	no terms about custody
3	**Work**	"employee" classification	minimum health insurance coverage

Example 1 (Commerce). *The Uniform Commercial Code and "Reasonable" Defaults.*

We can use the results of this section to understand aspects of the Uniform Commercial Code. Consider first the issue of contracting on delivery times. Suppose that θ takes on only two states, leading to a high or a low surplus, respectively. Section 3.2 showed that without a default rule, the "low" realization of θ may be so bad for the seller that the seller would avoid entering the contract to begin with. In other words, the distribution of surplus in this state is such that the seller would not enter. In this scenario, the government has an "equity" objective ($\beta > 0$), to smooth the distribution of surplus, which is based in a goal of providing efficient investment incentives. Proposition 2 suggests that when the government has some "equity" objective $\beta > 0$, it can implement a "reasonable" default rule

for delivery times which achieves the first best distribution of surplus (and therefore the efficient contract) in both states of the world $\{\theta_l, \theta_h\}$.

The U.C.C. also has some immutable rules, which place limits on the kinds of contracts that will be enforced in court. For instance, the law prevents contract terms stipulating times deemed "manifestly unreasonable." If times for delivery are too extreme, there will likely be the need for arbitration. Arbitration can be costly and therefore strain public resources. In other words, there is an externality that arises out of contracts with extreme times for delivery, which is mitigated with the "manifestly unreasonable" immutable rule.

Example 2 (Marriage). *"Equitable Division" Defaults and Limited Enforcement.*

In the regulation of marriage contracts in the U.S., the government supplies default marriage rules that govern unless expressly contracted around. For instance, if to-be-spouses get married without signing a prenuptial agreement that specifies otherwise, the division of assets upon divorce will follow an "equitable division" rule in many states. The default rule fills a gap in an incomplete contract, and ensures a particular distribution of surplus in the event of divorce. In this case, unlike in Example 1, the government may have preferences directly over the "equity" of a contract, for fairness reasons (Bix, 1998). That is $\beta > 0$ because the government wants to avoid unjust outcomes. Proposition 1 thus helps to illustrate why this equitable division rule exists: it serves as a default that allows the regulator to achieve first best.

Proposition 3 can help to understand the limits on the contracts the government is willing to enforce. The best allocation of custody from the perspective of the government is the allocation that is best for the child. This allocation may not align with what the spouses view to be the best allocation. Suppose the government has an infinitely negative payoff when full custody is given to the "wrong" spouse. Then, our model predicts that the government would not enforce *any* contract terms involving custody of children, child support, or visitation. In line with this prediction of our model, there is no state in the U.S.

which will enforce terms about children in prenuptial contracts.[12]

Example 3 (Work). *The Classification of Platform Workers*.

We return to the case of regulating app-based platform work in detail in section 3.5. For now, it is sufficient to establish the premise that to the extent that the regulator is able to affect the default, they will choose to shift the default contract in order to push the resulting distribution of outcomes across states closer to their preferred distribution. Furthermore, the presence of externalities justifies the restriction of the contracting space inherent in mandating minimum health and unemployment insurance coverage.

One aspect of this setting that makes interpretation more subtle is that there are multiple policy levers through which the government influences “default" labor contracts, i.e. the labor contracts that serve as an outside option in specific employee-employer negotiations. (This subtlety contrasts with the incomplete contracting setting where the regulator directly establishes defaults which hold in case they are not explicitly contracted away.) A primary avenue through which governments affect worker “defaults" is worker classifications, which establish that certain forms of work in certain sectors must be governed by “employee" rules and not “independent contractor" rules. Regulators also use the threat of legislation and other forms of soft power to shift the default (Stewart and Stanford, 2017). For instance, Bernie Sanders’ Stop BEZOS Act may be partially responsible for Amazon’s subsequent decisions to raise wages.[13]

In the platform work example, the unverifiable state of the world θ is the degree to which workers value benefits relative to cash. Direct surveys show that there is a large degree of heterogeneity in θ (e.g. Gruber (2022)) justifying our continuous treatment of Θ. We use the platform work example to characterize the optimal default delegation policy. Comparative statics on the optimal policy help us understand recent and anticipated shifts in the approach to regulating platform work.

---

[12] https://www.findlaw.com/family/marriage/what-can-and-cannot-be-included

[13] See, for example, Bhattarai (2018).

3.5 Second Best Analysis: An Application to the Regulation of Platform Work

To summarize, the prior section showed that default delegation attains any implementable first-best contract when the regulator has only equity concerns (Proposition 2), and can attain the first-best contract in the presence of both equity and externality concerns in some special cases (Proposition 5). Although a range of settings can be reduced to a two-state case (see section C.3), the limitations detailed above suggest that developing intuition for regulation in practice requires understanding optimal default delegation when it *does not* attain the first best, i.e. when the conditions in Proposition 6 do not hold.

In this section, we characterize the expected welfare-maximizing default delegation policy via comparative statics and closed-form solutions for a specific example. We find that the intuition developed in section 3.4 helps to understand the trade-offs a regulator faces when the first best is unattainable with default delegation: even when it does not achieve first best, default delegation balances the distribution of surplus across states (to mitigate inequities between parties) and internalizes externalities. Comparative statics show how the welfare-maximizing default delegation policy $\{\mathcal{Q}, d\}$ changes with: (1) the regulator's preferences over equity (β) and externalities (γ), (2) the regulator's prior about the unknown state of the world, and; (3) differences in the agents' bargaining weights.

This section also doubles as an in-depth investigation into the regulation of a rapidly evolving form of labor contract. We use the theory of default delegation to generate predictions about the regulation of platform work.

3.5.1 Set up

Consider a regulator concerned about the contracts written between a continuum of firms and their app-based workers. The regulator is uncertain about how any given worker values benefits q relative to cash c. To fix ideas, let q be the demand for health insurance which varies in terms of its coverage. We capture the regulator's uncertainty by θ, an unknown

state of the world drawn independently from a distribution $G(\theta)$ with support $[\underline{\theta}, \overline{\theta}]$.

Although θ is unknown to the regulator, it is common knowledge to each worker-firm pair. Each worker's preference is quadratic and reaches a maximum at $q = \theta$. Some workers put relatively high value on additional worker benefits relative to additional pay while others receive relatively little additional surplus from additional benefits.[14]

The firm's costs are state-independent: q^2 is the known cost of providing benefits. These costs are convex, and the firm always prefers to provide the minimum level of coverage. To summarize, workers and the firm have preferences

$$\textbf{Firm: } U_f(q, c) = \Pi = R - q^2 - c \quad \textbf{Worker: } U_w(q, c; \theta) = \text{WS} = w - (q - \theta)^2 + c$$

where w refers to an exogenous worker wage and R is the firm's revenue net of this wage.

We model the externalities generated by the contract as linear in q, for ease of exposition. (The results presented in this section would be qualitatively similar for increasing functions of q.) Since the externality term is linear, the socially optimal level of q is always above the agent-optimal quality level. This externality term captures the fact that the regulator would be forced to cover health care that is not covered in the employer provided health insurance.

$$\textbf{Externality: } U_r(q) = \gamma q$$

All of these terms enter into the SWF presented in section 3.3.

$$\text{SWF} = \underbrace{\text{WS} + \Pi}_{\text{"efficiency"}} \underbrace{-\beta(\text{WS} - \Pi)^2}_{\text{"equity"}} + \underbrace{\gamma q}_{\text{"externality"}}$$

The regulator's program. In a second-best environment, the regulator aims to maximize *expected* welfare. The regulator solves for the expected social welfare-maximizing default delegation policy by solving the following program:

$$\max_{\{\mathcal{Q}, (q_d, c_d)\}} \mathbb{E}_G[SWF(q_\theta, c_\theta; \theta)] \quad s.t. \quad q_d \in \mathcal{Q}$$

[14] See Gruber (2022) for survey data reporting substantial heterogeneity in gig-worker's valuations of benefits.

where

$$(q_\theta, c_\theta) = \underset{q \in \mathcal{Q}, c \in \mathbb{R}}{\arg\max} (WS(q, c; \theta) - WS(q_d, c_d; \theta))^\delta (\Pi(q, c) - \Pi(q_d, c_d))^{1-\delta}.$$

That is, the regulator chooses a delegation set $\mathcal{Q}$ and default q_d, c_d to maximize expected welfare with respect to its prior $G(\theta)$, foreseeing that the agents will Nash bargain. The default quality level must lie in the delegation set ($q_d \in \mathcal{Q}$).

In order to solve the regulator's program, we make use of results from single-agent delegation problems. Note that conditional on a particular default, the regulator interacts with the agents as if they were a single combined agent (maximizing their joint surplus). Alonso and Matouschek (2008) characterizes settings in which the optimal delegation set for a single agent is an *interval*. Building on their results, we restrict attention to cases in which $\mathcal{Q}$ is an interval, i.e. $\mathcal{Q} = [\underline{q}, \bar{q}]$.[15]

Let q_θ be the quality level that the agents choose conditional on $\{\mathcal{Q}, d\}$ when the state of the world is θ. Note that $\frac{\theta}{2}$ is the value of q that maximizes the joint surplus of the workers and the firm. The agents will bargain to the value in the delegation interval that maximizes their joint surplus. If the joint surplus-maximizing value of q is not in the delegation interval, then q_θ will be an endpoint of the interval.

Using the fact that the optimal delegation set is an interval, and assuming that the worker and firm bargain to the value q_θ in the delegation set that maximizes their joint surplus, we can rewrite the regulator program:

$$\max_{\{c_d, q_d, \underline{q}, \bar{q}\}} \int_{\underline{\theta}}^{\bar{\theta}} \left(WS(c_\theta, q_\theta; \theta) + \Pi(c_\theta, q_\theta) + \gamma q_\theta - \beta \left(WS(c_\theta, q_\theta; \theta) - \Pi(c_\theta, q_\theta) \right)^2 \right) dG(\theta)$$

[15] We discuss the use of interval delegation in more detail in Appendix C.5.1 and show conditions under which the optimal default delegation mechanism will take the form of a closed set.

subject to

$$(q_\theta, c_\theta) = \begin{cases} (\underline{q}, c(\underline{q}, c_d, q_d; \theta)) & \theta < 2\underline{q} \\ (\frac{\theta}{2}, c(\frac{\theta}{2}, c_d, q_d; \theta)) & 2\underline{q} \leq \theta \leq 2\overline{q} \\ (\overline{q}, c(\overline{q}, c_d, q_d; \theta)) & \theta > 2\overline{q} \end{cases}$$

where c_θ is an implicitly defined function

$$c_\theta = c(q_\theta, c_d, q_d; \theta) = c_d + (1-\delta)(-(q_d - \theta)^2 + (q_\theta - \theta)^2)) - \delta(-q_d^2 + q_\theta^2)). \tag{3.14}$$

In the remainder of this section, we solve the regulator's program in three different bargaining regimes. As the regulator's program is sensitive to the exogenous bargaining parameter δ, we simplify the analysis by considering three cases: equal bargaining ($\delta = .5$), firm control ($\delta = 0$) and worker control ($\delta = 1$). In order to get explicit solutions, we will sometimes make specific assumptions about the distribution of types (e.g. that the regulator's prior is uniformly distributed on the unit interval, i.e. $G(\theta) = \text{Unif}[0, 1]$).

3.5.2 Firm control

We begin our second-best analysis with the case in which the workers have no bargaining power, i.e. $\delta = 0$. In this case, the firm makes take-it-or-leave-it contracts to the workers. This assumption on δ aligns best with the facts about gig work in the U.S. Since gig-workers are not classified as employees, they are not protected by collective bargaining laws and so cannot form unions. Furthermore, factors that are not specific to gig-work contribute to low worker bargaining power in the U.S.: declining unionization (Farber *et al.*, 2021), rising employer concentration (Benmelech *et al.*, 2020), and diminished worker protections all contribute to low levels of worker bargaining power across sectors (Summers and Stansbury, 2020), even as superstar firms capture exceptional profits (Autor *et al.*, 2020a).

Solving the regulator problem. When $\delta = 0$, the welfare-maximizing default delegation policy is complex. In particular, the endpoints of the delegation set $\underline{q}$ and $\overline{q}$ as well as the

default q_d, c_d are all interdependent. Each is a function of the other terms, as well as the equity and externality parameters β and γ.

As a baseline, it is useful to begin our analysis with a simple closed-form solution that obtains under further assumptions. We assume that types are distributed uniformly on the unit interval.

Assumption 4 *Types θ are drawn from the uniform distribution, $\theta \sim \text{Unif}[0,1]$.*

We first consider a case in which the maximum $\overline{q}$ and minimum $\underline{q}$ do not constrain bargaining in any state of the world.[16]

Assumption 5 *The outcome of bargaining q_θ is $\frac{\theta}{2}$ in all states.*

Under Assumptions 4 and 5, we get the following expected welfare-maximizing default:

$$q_d^* = \frac{3}{8}, \quad c_d^* = \frac{R-w}{2} + \frac{1}{64}. \tag{3.15}$$

The default transfer is above the level which equates the exogenous components of surplus, $\frac{R-w}{2}$. Similarly, the default quality level is closer to the worker's expected optimal quality level, $q = \frac{1}{2}$, then the firm's optimal, $q = 0$. This default illustrates the following, more general observation.

Result 1 *When the firm has all the bargaining power ($\delta = 0$), the welfare-maximizing default (q_d^*, c_d^*) favors the worker in terms of both the quality level and the transfer.*

To see why this result holds more generally, recall that $\delta = 0$ implies that the firm receives all of the surplus from bargaining. So, in order to equalize the surplus between the two parties, the equity-concerned regulator uses the default to improve the worker's position before bargaining occurs.

But the optimal q_d^* does not deliver an exactly equal split of the surplus—instead, it *minimizes* the inequity across all possible states. The leftmost graph in Figure 3.3 plots the worker's utility, firm's profits, and the inequity term in the social welfare function (squared

[16] This occurs when β and γ are both small.

difference between worker profit and firm profit) evaluated at q_d^*, as a function of the state θ. The worker utility is $u(q_d^*; \theta)$, i.e. their value of the default across states. The firm's profit is their value of the default plus the bargaining surplus. In this case the observed inequity is minimized across states.

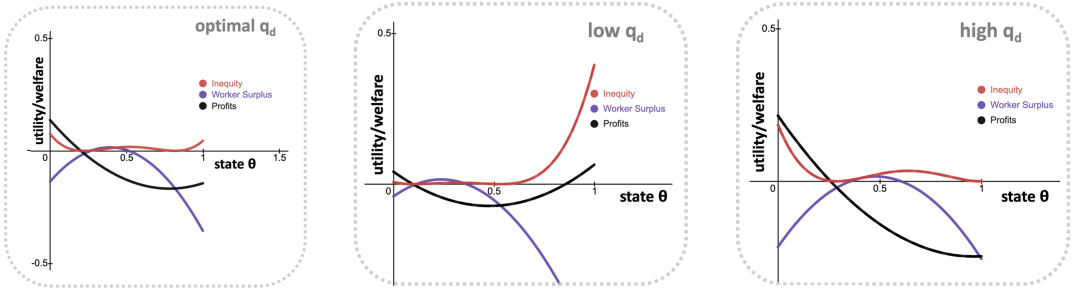


Figure 3.3: *Distribution of Surplus Across States with Varying Default Quality*

To see more clearly how inequity is affected by the default, the middle and right graphs in Figure 3.3 show how decreasing q_d above or below the optimal level exacerbate inequity. In particular, a lower default quality level $q_d < q_d^*$ exacerbates inequity in the high states while mitigating it in the low states (middle). A higher default quality level $q_d > q_d^*$ exacerbates inequity in low states while mitigating it in the high states (right).

Minimums mitigate externalities and influence the default

So far we have focused exclusively on the optimal default (q_d^*, c_d^*), since under Assumption 5, the delegation set does not constrain the agents. How does the optimal default change as the regulator constrains the delegation set $\mathcal{Q}$, i.e. as the regulator raises the minimum $(\underline{q})$ or lowers the maximum $(\overline{q})$?

Recall from the results in section 3.4 that the main reason the regulator constrains the delegation interval is to internalize externalities. In this example, since there is a positive externality associated with increased benefits, the regulator imposes a minimum quality level when the value of the externality is large. This, in turn, influences the optimal default. When Assumption 5 does not hold, the optimal default quality is a function of the minimum

and maximum:

$$q_d^* = \frac{1 + 2\overline{q} - 8(1-\overline{q})\overline{q}^3 + 8(1-\underline{q})\underline{q}^3}{4}. \tag{3.16}$$

The optimal default is increasing in the minimum $\underline{q}$ and decreasing in the maximum $\overline{q}$. This is because raising the minimum reduces the extent to which an increase in the default leads to higher inequity in the low states.

Result 2 *The default quality level q_d^* is increasing in the minimum quality level $\underline{q}$ and decreasing in the maximum quality level $\overline{q}$.*

Discussion of Results 1 and 2. The first two results help us understand how regulators with preferences over efficiency, equity and externalities choose regulatory policies when one party has all the bargaining power.

In the context of regulating the platform worker contracts in the U.S., the assumption that firms have all the bargaining power is a natural one in light of the fact that workers are not protected under collective bargaining laws. Result 1 suggests that an equity-concerned regulator would set a default level of benefits q_d and default wages c_d to be closer to the worker's optimum than to the firm's optimum.

As we've discussed, regulators may not be able to influence the default directly. Unlike in standard incomplete contracting environments, the default in an employment contract is a worker's outside option—this default will be influenced by labor laws but also by industry characteristics that may be outside the regulator's control. However, the default is likely to be an increasing function of the minimum level of benefits $\underline{q}$. This implies that a regulator attempting to internalize externalities and increase the default quality level can do so by raising $\underline{q}$. Furthermore, Result 2 suggests that these measures are not opposed but rather reinforcing. Raising the minimum benefit level would further raise the optimal default. Together, these results suggest that raising the enforced minimum benefits provision in platform work could serve two purposes: it helps the regulator internalize externalities (shifting social insurance costs onto the firm) while also helping the regulator achieve equity goals (by shifting the default upward).

In the U.S., platform workers were initially classified as contractors and not employees. As contractors, the default favors the firm as there are no enforced minimums: firms are neither expected nor required to provide benefits. This fact suggests that initially, regulators may not have been concerned about the inequities and externalities that may result from the contract. To summarize:

- Result 1 predicts that equity-concerned regulators set the default level of benefits provision and pay to be closer to the worker's desired level than the firm's desired level.
- If the regulator cannot influence the default directly, it can increase the minimum benefits level. In such a case, the minimum serves to both internalize externalities *and* equalize surplus between the parties.

How changes in the social welfare function affect default delegation

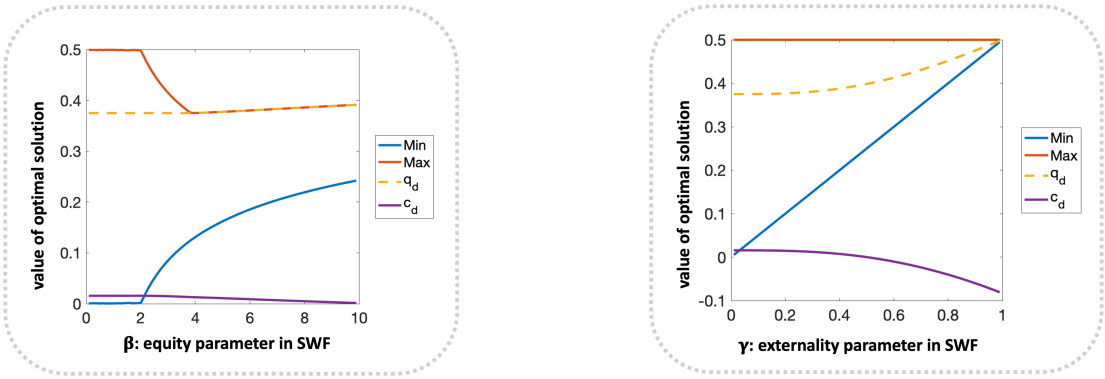
Results 1 and 2 give insight into how the optimal default (q_d^*, c_d^*) is set when the firm has all the bargaining power. These results help us explain how a regulator with equity and externality concerns maximizes social welfare. We next look at how the expected welfare-maximizing default delegation policy changes with the parameters in the regulator's social welfare function.

Governments' preferences over equity and externalities are not fixed. The social cost of inequities between different parties, captured by β, changes over time as social preferences shift and as different political parties transition into and out of power. Externalities from a particular kind of contract also change as the composition of people and firms entering that contract shift over time. Our analysis of optimal default delegation helps us understand how regulatory policy reacts to such shifts.

We first consider how the optimal default delegation policy changes with the size of the inequity penalty β. We continue to assume that types are uniform (Assumption 4 holds). Panel (a) of Figure 3.4 plots the optimal default delegation policy $(\mathcal{Q}^*, d^*)$ as a function of the inequity penalty parameter (β), assuming there are no externality concerns ($\gamma = 0$) to

isolate the effect of β. When $\gamma = 0$, the optimal delegation interval $\mathcal{Q}^* = [\underline{q}^*, \overline{q}^*]$ simply trades off losses in efficiency for gains in equity. Figure 3.4 panel (a) shows that at low levels of β ((precisely: $\beta \in (0, 2)$), the efficiency concerns dominate equity concerns, and the regulator does not constrain the interval ($[\underline{q}, \overline{q}]$). Note that on this interval $\beta \in (0, 2)$ the optimal default q_d is constant at the value in (3.15) because Assumption 5 holds, and as soon as β is strictly positive, the regulator's first-best has an equal distribution of surplus. We call the point beyond which equity concerns dominate efficiency concerns ($\beta = 2$ in this example) the *equity-efficiency threshold*.

When β passes the equity-efficiency threshold, further increases in β shrink the optimal delegation interval: the maximum $\overline{q}^*$ decreases and the minimum $\underline{q}^*$ increases.[17] This is because the bargaining is unequal ($\delta = 0$), so if inequity is very costly to the regulator, it wants to limit the degree to which parties can bargain. It is somewhat counterintuitive that the regulator would want to decrease the maximum since higher benefits q make the worker better off. Recall that all of the gains from bargaining are accrued by the firm, leading to high levels of inequity as we have defined it. Eventually, the maximum converges to the default and further improvements in equity are obtained by increasing $\overline{q}^*$, q_d^* and $\underline{q}^*$.



(a) $\delta = 0, \gamma = 0$ **(b)** $\delta = 0, \beta = 0$

Figure 3.4: *Relationship Between Externality and Equity on Optimal Default Delegation*

[17] This is one of main differences between the first-best and second-best analysis: in the first-best analysis, Proposition 2 shows that the regulator cannot do better from an equity perspective by constraining the set of enforced contracts. However, as most cases of interest will feature levels of β that are below the equity-efficiency threshold, the intuition that the regulator constrains the interval mainly to curb externalities holds.

Result 3 *When the firm has full bargaining power ($\delta = 0$), as equity (β) concerns increase,*

(i) up to the equity-efficiency threshold, the optimal delegation interval $[\underline{q}^*, \overline{q}^*]$ and default quality q_d^* are constant, and

(ii) beyond the equity-efficiency threshold, $[\underline{q}^*, \overline{q}^*]$ shrinks to reduce the unequal gains from bargaining. The default quality q_d^* responds following Result 2, and the default transfer c_d^* compensates this adjustment.

Panel (b) in Figure 3.4 shows how changes in γ affect the optimal default delegation policy assuming (to isolate the effect of γ) that there are no equity concerns ($\beta = 0$). As the externality term γ increases, the minimum $\underline{q}$ increases. This straightforwardly highlights the primary rationale for limiting the interval of enforced contracts as introduced in Proposition 4: restricting the interval of enforced contracts mitigates externalities. The default quality level q_d^* is also increasing in γ through $\underline{q}^*$ (this is solely through the channel presented in Result 2, which tells us that q_d^* is increasing in $\underline{q}^*$).

Result 4 *When the firm has full bargaining power ($\delta = 0$), as externality concerns (γ) increase, the optimal default-delegation policy restricts the delegation interval $[\underline{q}^*, \overline{q}^*]$ in the direction which internalizes the externality. The default quality q_d^* responds to the change in $\underline{q}^*$ or $\overline{q}^*$ following Result 2, and the default transfer c_d^* compensates this adjustment.*

Discussion of Results 3 and 4. Results 1 and 2 suggest that regulators set defaults and minimums to favor workers in the presence of externality and equity concerns. The fact that initially platform workers were classified as contractors in the U.S. is out of step with this prediction, and suggests that regulators may *not* have been concerned about externalities and inequities resulting from platform labor contracts, at least initially. The results of this subsection help us understand one reason why discussions about the classification of platform workers (which affects both enforced minimums and the default contract) are now widespread in the U.S. and other markets: the social welfare function may have shifted.

Result 3 shows that when the inequity penalty is below the equity-efficiency threshold, increases in the inequity penalty parameter do not affect the optimal default-delegation policy. Beyond the equity-efficiency threshold, Result 3 part (ii) shows that increases in the inequity penalty β begin to constrain the delegation interval to avoid unequal bargaining. There is no reason to believe that the extreme case described in part (ii) is relevant to the regulation of platform work in the U.S.: when β is beyond the equity-efficiency threshold, the regulator is willing to burn surplus in order to improve equity outcomes. Rather it is more likely that a shift has occurred in the externalities or (regulator perceptions of) average preferences of workers, which we discuss below.

The externalities from gig work contracts may also be shifting. Evidence suggests that as platform work has gained popularity, the share of platform workers who treat platform work as a full time job has increased. This suggests that more gig workers do not have benefits provided by another employer, and so may rely on public benefits (ADP Research Institute (2020)). In our model, this suggests an increase in the externality term γ. Result 4 predicts that when this term increases, the optimal minimum increases. This prediction thus explains one reason for passing a bill such as AB5. To summarize:

- Result 3 part (i) suggests that when regulators have non-zero equity concerns, small changes in the inequity penalty would not affect the regulatory approach to platform work.
- Result 4 suggests that as externalities from platform work contracts increase, regulators increase enforced minimums of benefits provision $\underline{q}^*$.

3.5.3 Equal bargaining

We next turn to the special case in which the worker and the firm have equal bargaining power, i.e. $\delta = \frac{1}{2}$. This is a special case in which the allocation of bargaining power is *aligned* with the regulator's equity objective (to split the surplus equally). More generally, we say that bargaining power is *aligned* with social welfare when the regulator prefers to give the

worker some $\alpha \in (0,1)$ share of the surplus, and the worker's bargaining power is $\delta = \alpha$. In such cases, bargaining does not have the potential to exacerbate inequity.

Default delegation simplifies under aligned bargaining

As a result of the alignment of bargaining power and social objectives, the optimal default-delegation policy dramatically simplifies. The default (q_d^*, c_d^*) plays a single role: it addresses the expected inequity between the parties. The delegation interval also plays a single role: it mitigates the externalities. We characterize the optimal default delegation policy in Proposition 7.

Proposition 7 *Suppose the regulator's social welfare function is*

$$U_w + U_f - \beta(\alpha U_f - (1-\alpha)U_w)^2 + \gamma U_r(q) \tag{3.17}$$

where α represents the worker share of total surplus that is socially optimal. If $\delta = \alpha$ then

(i) $\underline{q}^*$ and $\overline{q}^*$ do not depend on q_d^* or c_d^*, and

(ii) q_d^* and c_d^* do not depend on γ, β, $\underline{q}^*$ or $\overline{q}^*$.

To illustrate this case more fully, we return to our example and solve for the optimal default-delegation policy under the uniform assumption (Assumption 4). The optimal default-delegation policy is given by

$$q_d^* = \frac{1}{2} \quad c_d^* = \frac{R-w}{2} - \frac{1}{12} \quad \underline{q}^* = \frac{\gamma}{2}. \tag{3.18}$$

The default quality q_d^* is equal to the worker's ex-ante expected optimal quality level,

$$\mu_\theta \equiv \int_{\underline{\theta}}^{\overline{\theta}} \theta dG(\theta).$$

In this case, it is efficient for the regulator to "pay" the workers in kind by setting the default quality level q_d^* to their expected optimal μ_θ. The default transfer c_d^* instead favors the firm. This arrangement is optimal in this case because the workers have state-dependent preferences whereas the firm does not. The regulator minimizes the expected losses due to

sub-optimal quality for the worker and equalizes the surplus through transfers to the firm. The minimum $\underline{q}^*$ is strictly a function of the externality term. Its slope is determined by equating the loss of efficiency to the gains coming from mitigating the externality.

Discussion of Proposition 7. When the allocation of bargaining power is aligned with the regulator's desired distribution of surplus, the optimal default-delegation policy is simpler. This result helps us understand two aspects of the regulation of platform work.

First, recall that what is special about this case is *not* that the bargaining parameter takes on a particular value, but that the bargaining parameter *is aligned with the regulator's equity objective*. The key insight Proposition 7 raises in the context of platform work is that when bargaining is aligned and the regulator can set an optimal default, the rationale for increasing minimum benefits provision in platform employment contracts is purely based in mitigating externalities. Comparing this result to Result 2 shows that as soon as the bargaining parameters are misaligned with social objectives, there is an additional, equity-based rationale for increasing minimums (when $\delta < \alpha$).

Second, the result on aligned bargaining raises a provocative question: what if the regulator could also choose the bargaining parameter? The comparison of these two cases show that the regulator would have significant incentive to adjust the bargaining parameter such that it is aligned with the regulator's objective. To summarize:

- Proposition 7 shows that when bargaining power is aligned with equity objectives ($\delta = \alpha$), the regulator sets minimums only to mitigate externalities.
- Proposition 7 is an analogue of the "countervailing power" argument (Galbraith, 1952): it suggests that the regulator's role is much more straightforward when there are institutions such as labor unions that align the share of bargaining power with social objectives.
- Proposition 7 may explain why some advocacy groups (who are not lawmakers and so cannot directly set the minimums) are seeking alternative means of building worker

power given that workers are not currently protected under collective bargaining laws.[18]

How changes in the regulator's prior affect default delegation

We next take advantage of the simplicity of the aligned bargaining case to study how shifts in the regulator's prior influence the optimal default-delegation policy. In particular, we compare the optimal default-delegation policy in (3.18) under the uniform distribution (Assumption 4) to a distribution with higher density in higher states.

Assumption 6 *Types θ are drawn from the distribution $G(\theta) = \theta^2$.*

Under Assumption 6, the optimal default-delegation policy is given by

$$q_d^* = \frac{3}{5} \quad c_d^* = \frac{R-w}{2} - \frac{3}{20} \quad \underline{q}^* = \frac{3\gamma}{4}. \tag{3.19}$$

Comparing (3.18) and (3.19) is instructive. As the cumulative distribution function of θ shifts towards higher states, the default quality q_d^* increases. However, the default quality level does not increase one for one with the expected state μ_θ. In the uniform case, (3.18) shows that $q_d^* = \mu_\theta = \frac{1}{2}$. With $G(\theta) = \theta^2$, (3.19) shows that $q_d^* = \frac{3}{5}$ whereas $\mu_\theta = \frac{2}{3}$. In the quadratic case, $\mu_\theta > q_d^*$ because the cost of raising the default quality q_d^* is relatively high in the low states. Even so, the higher default quality level q_d^* imposes a greater cost on the firm, which must be offset through the default transfer c_d^*: the transfer from the worker to the firm grows from $\frac{1}{12}$ in the uniform case to $\frac{3}{20}$ in the quadratic case.

Meanwhile, the minimum $\underline{q}^*$ in the quadratic case is steeper in γ than in the uniform case. This is because the efficiency cost of raising the minimum is reduced due to the lower density in the low states.

Result 5 *Let $G'(\theta)$ be a distribution that first-order stochastically dominates $G(\theta)$. The optimal default quality level q_d^* and the optimal minimum quality level minimum quality level $\underline{q}^*$ are higher under $G'(\theta)$ than under $G(\theta)$, and the optimal default transfer c_d^* is lower.*

[18] See, for example,
https://www.vox.com/2019/5/8/18535367/uber-drivers-strike-2019-cities

Discussion of Result 5. When the regulator's prior about worker types shifts upward, both the optimal default and minimum benefits level shift upward as well. In the case of platform work, the unknown state of the world is how workers value benefits relative to additional income. This effect suggests another explanation for why regulators may attempt to reclassify platform workers.

- Result 5 suggests that if more workers come to value benefits more highly (or if the regulator believes this to be the case) then the regulator would optimally increase the minimum level of benefits and increase the default.
- The effect of an upward-shifted prior, described in Result 5, has a similar result to a shift in the externality parameter, described in Result 4. The two hypothesized shifts are related: as more platform workers value benefits more highly, this is likely to come along with fewer platform workers getting benefits from other sources (and thus higher externalities). In both cases, the regulator would like to raise the minimum (and the default). In the case of Result 5, this effect occurs because more workers valuing benefits more highly makes the efficiency-externality trade-off more favorable (higher $\underline{q}$ is less costly from an efficiency perspective). In the case of Result 4, this effect occurs because the value of mitigating the externality is simply "worth more" to the regulator, in terms of efficiency.

3.5.4 Worker control

Finally, we consider the case in which workers have all of the bargaining power ($\delta = 1$). Although this case does not align with the facts about platform work in any context we know of, it is a theoretically valuable case. In particular, it highlights the role that state-dependent preferences are playing in the preceding analysis.

Proposition 8 *Assume that only one agent i has state-dependent preferences, and that this agent has all of the bargaining power (i.e. $\delta_i = 1$). Then,*

(i) the optimal default q_d^*, c_d^* is not uniquely determined, and

(ii) the expected social welfare under the optimal default-delegation policy is strictly lower when $\delta_i = 1$ *than when* $\delta_i = 0$.

Proposition 8 stems from the fact that the regulator has limited ability to distribute surplus across states when the agent with state-independent preferences has no bargaining power. So, any attempt to improve equity necessitates a greater efficiency trade-off. As a result, the expected welfare from the optimal default delegation policy is lower when the agent with state-independent preferences has no bargaining power.

In the context of regulating new labor contracts, Proposition 8 is somewhat counter-intuitive. It implies that regulatory solutions that place all of the bargaining power on the workers—to the point where workers are making take-it-or-leave-it offers to the firm—are unappealing from the perspective of an equity-concerned regulator who would like to achieve a particular distribution of surplus between the worker and the firm assuming an "equal split" objective. The optimal solution when workers have all the bargaining power ($\delta = 1$) leads to strictly lower social welfare than the optimal solution when workers have no bargaining power ($\delta = 0$).

Proposition 8 is related to Corollary 2 in section 3.4, which highlights that the set of implementable outcomes is larger when the party that is less sensitive to the state has more bargaining power. In both cases, the regulator loses some of its ability to influence the state-dependent distribution of surplus when the party with more bargaining power is less sensitive to the state. In other words, when bargaining power shifts away from alignment and toward the agent with more sensitive preferences, the regulator's trade-off between discretion and control is exacerbated.

3.6 Conclusion

This paper makes two main contributions. First, we provide mechanism design foundations for the widespread use of default and immutable rules in contract law. Governments primarily use immutable rules to mitigate externalities while they use default rules to affect

the distribution of surplus between contracting parties. We argue that the logic behind default and immutable rules applies directly in the case of incomplete contracts but can also be applied more generally to the regulation of bilateral contracts whenever the government's social welfare function weights efficiency, equity and externalities to some degree. We characterize circumstances under which *default delegation* achieves first-best in terms of the government's tradeoffs among efficiency, distributional concerns, and externalities.

Second, we show how the government's optimal default delegation policy depends on underlying parameters of the contracting environment as well as its social welfare function. We apply our results to the regulation of platform work, a vast and growing labor arrangement in the U.S. and abroad. Our results organize the debate around the classification of app-based platform workers.

Overall, this paper has provided a theoretical lens for the design of bilateral contracting environments and its effects on social welfare. In future theoretical work, we hope to use this framework to directly analyze the tradeoffs in particular organizational contexts in which a principal delegates a decision to bargaining agents. On the empirical side, we hope to use this framework to explain observed heterogeneity in government regulation.

References

ADP Research Institute (2020). *Illuminating the Shadow Workforce: Insights Into the Gig Workforce in Businesses*. Tech. rep.

Aghion, P., Bergeaud, A., Boppart, T., Klenow, P. J. and Li, H. (2019). A Theory of Falling Growth and Rising Rents. (26448).

—, Dewatripont, M. and Rey, P. (1994). Renegotiation design with unverifiable information. *Econometrica: Journal of the Econometric Society*, pp. 257–282.

— and Howitt, P. (1992). A Model of Growth through Creative Destruction. *Econometrica*, **60** (2), 323–51.

Alonso, R. and Matouschek, N. (2008). Optimal delegation. *Review of Economic Studies*, **75** (1), 259–293.

Altonji, J., Arcidiacono, P. and Maurel, A. (2016). The analysis of field choice in college and graduate school. In *Handbook of the Economics of Education*, vol. 5, Elsevier B.V, pp. 305–396.

Altonji, J. G., Blom, E. and Meghir, C. (2012). Heterogeneity in human capital investments: High school curriculum, college major, and careers. *Annual review of economics*, **4** (1), 185–223.

Andrews, R. J., Imberman, S. A., Lovenheim, M. F. and Stange, K. M. (2022). *The Returns to College Major Choice: Average and Distributional Effects, Career Trajectories, and Earnings Variability*. Working Paper 30331, National Bureau of Economic Research.

Atkeson, A. and Burstein, A. (2008). Pricing-to-Market, Trade Costs, and International Relative Prices. *American Economic Review*, **98** (5), 1998–2031.

Autor, D., Dorn, D., Katz, L. F., Patterson, C. and Van Reenen, J. (2020a). The Fall of the Labor Share and the Rise of Superstar Firms. *The Quarterly Journal of Economics*.

—, Goldin, C. and Katz, L. F. (2020b). Extending the race between education and technology. *AEA papers and proceedings*, **110**, 347–351.

Autor, D. H. (2019). Work of the past, work of the future. *AEA Papers and Proceedings*, **109**, 1–32.

—, Katz, L. F. and Kearney, M. S. (2006). The polarization of the u.s. labor market. *The American economic review*, **96** (2), 189–194.

Ayres, I. and Gertner, R. (1989). Filling gaps in incomplete contracts: An economic theory of default rules. *The Yale law journal*, **99** (1), 87–130.

Baron, D. P. and Myerson, R. B. (1982). Regulating a monopolist with unknown costs. *Econometrica: Journal of the Econometric Society*, pp. 911–930.

Bell, A., Chetty, R., Jaravel, X., Petkova, N. and Van Reenen, J. (2019). Who becomes an inventor in america? the importance of exposure to innovation. *The Quarterly journal of economics*, **134** (2), 647–713.

Benmelech, E., Bergman, N. K. and Kim, H. (2020). Strong employers and weak employees: How does employer concentration affect wages? *Journal of Human Resources*, pp. 0119–10007R1.

Berry, C. R. and Glaeser, E. L. (2005). The divergence of human capital levels across cities. *Papers in regional science*, **84** (3), 407–444.

Bhattarai, A. (2018). Amazon boosts minimum wage to $15 for all workers following criticism. *Washington Post*, **October 2**.

Bix, B. (1998). Bargaining in the shadow of love: the enforcement of premarital agreements and how we think about marriage. *William and Mary law review*, **40** (1), 145.

Boudreau, K. and Marx, M. (2019). From theory to practice: Field experimental evidence on early exposure of engineering majors to professional work. *NBER Working Paper Series*, p. 26013.

Cao, D., Hyatt, H., Mukoyama, T. and Sager, E. (2019). Firm Growth Through New Establishments. *SSRN Electronic Journal*.

Cavenaile, L., Celik, M. A. and Tian, X. (2019). Are Markups Too High? Competition, Strategic Innovation, and Industry Dynamics. *SSRN Electronic Journal*.

Christiansen, C., Joensen, J. S. and Nielsen, H. S. (2007). The risk-return trade-off in human capital investment. *Labour economics*, **14** (6), 971–986.

Chung, T.-Y. (1991). Incomplete contracts, specific investments, and risk sharing. *The Review of Economic Studies*, **58** (5), 1031–1042.

Crossley, T., Gong, Y., Stinebrickner, T. R. and Stinebrickner, R. (2021). *Examining Income Expectations in the College and Early Post-college Periods: New Distributional Tests of Rational Expectations.* Working Paper 28353, National Bureau of Economic Research.

Davis, D. R. and Dingel, J. I. (2019). A spatial knowledge economy. *The American economic review*, **109** (1), 153–170.

De Loecker, J., Eeckhout, J. and Unger, G. (2020). The Rise of Market Power and the Macroeconomic Implications. *The Quarterly Journal of Economics*, **135** (2), 561–644.

Diamond, R. (2016). The determinants and welfare implications of us workers' diverging location choices by skill: 1980-2000. *The American economic review*, **106** (3), 479–524.

Dillon, E. W. (2018). Risk and return trade-offs in lifetime earnings. *Journal of labor economics*, **36** (4), 981–1021.

Easterbrook, F. H. and Fischel, D. R. (1982). Corporate control transactions. *The Yale law journal*, **91** (4), 698–737.

— and — (1989). The corporate contract. *Columbia Law Review*, **89**, 1416.

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

Edmond, C., Midrigan, V. and Yi Xu, D. (2019). *How Costly Are Markups?* mimeo.

Ellison, G., Glaeser, E. L. and Kerr, W. R. (2010). What causes industry agglomeration? evidence from coagglomeration patterns. *The American economic review*, **100** (3), 1195–1213.

Fajgelbaum, P. D. and Gaubert, C. (2020). Optimal spatial policies, geography, and sorting. *The Quarterly journal of economics*, **135** (2), 959–1036.

Farber, H. S., Herbst, D., Kuziemko, I. and Naidu, S. (2021). Unions and inequality over the twentieth century: New evidence from survey data. *The Quarterly Journal of Economics*, **136** (3), 1325–1385.

Galbraith, J. K. (1952). *American Capitalism: The Concept of Countervailing Power*. Routledge.

Glaeser, E. L. and Shleifer, A. (2001). A reason for quantity regulation. *The American economic review*, **91** (2), 431–435.

Glaeser, E. L. E. L. and Gottlieb, J. D. (2008). The economics of place-making policies. *Brookings papers on economic activity*, **2008** (1), 155–239.

Goetz, C. J. and Scott, R. E. (1983). The mitigation principle: toward a general theory of contractual obligation. *Virginia Law Review*, pp. 967–1024.

Green, J. R. and Laffont, J.-J. (1992). Renegotiation and the form of efficient contracts. *Annales d'Economie et de Statistique*, pp. 123–150.

Grossman, G. M. and Helpman, E. (1991). Quality Ladders in the Theory of Growth. *The Review of Economic Studies*, **58** (1), 43–61.

Gruber, J. (2022). *Designing Benefits for Platform Workers*. Working Paper 29736, National Bureau of Economic Research.

Gutierrez, G. and Philippon, T. (2019). *The Failure of Free Entry*. NBER Working Paper Series 26001, National Bureau of Economic Research, Inc.

Hart, O. and Moore, J. (1988). Incomplete contracts and renegotiation. *Econometrica: Journal of the Econometric Society*, pp. 755–785.

Hermalin, B. E. and Katz, M. L. (1991). Moral hazard and verifiability: The effects of renegotiation in agency. *Econometrica: Journal of the Econometric Society*, pp. 1735–1753.

Hersch, J. and Shinall, J. B. (2019). When equitable is not equal: experimental evidence on the division of marital assets in divorce. *Review of economics of the household*, **18** (3), 655–682.

Hillman, N. and Weichman, T. (2016). Education deserts: The continued significance of "place" in the twenty-first century. *Viewpoints: Voices from the Field. Washington, DC: American Council on Education.*

Holmström, B. (1977). "on incentives and control in organizations. *Ph.D. Thesis, Stanford University*.

Holmström, B. (1984). On the theory of delegation.

Hsieh, C.-T. and Rossi-Hansberg, E. (2020). *The Industrial Revolution in Services*. Tech. rep., Princeton.

Kantor, S. E. and Whalley, A. (2014). Knowledge spillovers from research universities: Evidence from endowment value shocks. *The review of economics and statistics*, **96** (1), 171–188.

Katz, L. F. and Murphy, K. M. (1992). Changes in relative wages, 1963–1987: Supply and demand factors. *The Quarterly journal of economics*, **107** (1), 35–78.

Kehrig, M. and Vincent, N. (2019). *Good Dispersion, Bad Dispersion*. NBER Working Paper Series 25923, National Bureau of Economic Research, Inc.

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

Kline, P. and Moretti, E. (2014). Local economic development, agglomeration economics, and the big push: 100 years of evidence from the tennessee valley authority. *The Quarterly journal of economics*, **129** (1), 275–332.

Krishna, V. and Morgan, J. (2001). A model of expertise. *The Quarterly Journal of Economics*, **116** (2), 747–775.

Laffont, J.-J. and Martimort, D. (1998). Collusion and delegation. *The Rand Journal of Economics*, pp. 280–305.

Long, M. C., Goldhaber, D. and Huntington-Klein, N. (2015). Do completed college majors respond to changes in wages? *Economics of education review*, **49** (December), 1–14.

Marshall, A. (1890). *Principles of economics*. Palgrave classics in economics, Palgrave Macmillan.

Martimort, D. and Semenov, A. (2008). The informational effects of competition and collusion in legislative politics. *Journal of Public Economics*, **92** (7), 1541–1563.

Maskin, E. (1999). Nash equilibrium and welfare optimality. *The Review of Economic Studies*, **66** (1), 23–38.

— and Moore, J. (1999). Implementation and renegotiation. *Review of Economic Studies*, pp. 39–56.

Molloy, R., Smith, C. L. and Wozniak, A. (2011). Internal migration in the united states. *The Journal of economic perspectives*, **25** (3), 173–196.

Mookherjee, D. (2006). Decentralization, hierarchies, and incentives: A mechanism design perspective. *Journal of Economic Literature*, **44** (2), 367–390.

Moore, J. and Repullo, R. (1988). Subgame perfect implementation. *Econometrica: Journal of the Econometric Society*, pp. 1191–1220.

Moretti, E. (2012). *The new geography of jobs*. Boston: Houghton Mifflin Harcourt.

— (2013). Real wage inequality. *American economic journal. Applied economics*, **5** (1), 65–103.

Parenti, M. (2018). Large and Small Firms in a Global Market: David vs. Goliath. *The Journal of International Economics*, **110**, 103–118.

Porter, C. and Serra, D. (2020). Gender differences in the choice of major: The importance of female role models. *American economic journal. Applied economics*, **12** (3), 226–254.

Posner, R. A. (1986). Economic analysis of law.

Romer, P. M. (1990). Endogenous Technological Change. *Journal of Political Economy*, **98** (5), S71–S102.

Rossi-Hansberg, E., Sarte, P.-D. and Schwartzman, F. (2019). Cognitive hubs and spatial redistribution. *Working paper (Federal Reserve Bank of Richmond)*, **19** (16), 1–83.

Rubinstein, A. and Wolinsky, A. (1992). Renegotiation-proof implementation and time preferences. *The American Economic Review*, pp. 600–614.

Schubert, G. (2021). House price contagion and u.s. city migration networks.

Shimomura, K.-I. and Thisse, J.-F. (2012). Competition Among the Big and the Small. *The RAND Journal of Economics*, **43** (2), 329–347.

Standard and Poors (S&P 2018a). *Compustat Daily Updates - Fundamentals Annual*. Tech. rep., Accessed from Wharton Research Data Services (WRDS): https://wrds-www.wharton.upenn.edu (Accessed on November 16, 2020).

Stange, K. M., Simon, A., Conzelmann, J. G., Martin, S. M., Hershbein, B. and Hemelt, S. W. (2022). Grads on the go: Measuring college-specific labor markets for graduates. *NBER Working Paper Series*.

Stewart, A. and Stanford, J. (2017). Regulating work in the gig economy: What are the options? *The economic and labour relations review : ELRR*, **28** (3), 420–437.

Summers, L. H. and Stansbury, A. (2020). The declining worker power hypothesis: An explanation for the recent evolution of the american economy.

Tirole, J. (1986). Hierarchies and bureaucracies: On the role of collusion in organizations. *Journal of Law and Economic Organization.*, **2**, 181.

Wiswall, M. and Zafar, B. (2015). Determinants of college major choice: Identification using an information experiment. *The Review of economic studies*, **82** (2 (291)), 791–824.

Zacchia, P. (2020). Knowledge spillovers through networks of scientists. *The Review of economic studies*, **87** (4), 1989–2018.

Appendix A

Appendix to Chapter 1

A.1 Comparative Dynamics

We first derive the slope of the $\dot{\lambda}_i$=0 curve. Differentiation of the right-hand side of (1.22) yields:

$$\hat{\Gamma}_i = -(\sigma-1)\,\hat{a}_i + \delta\hat{P} - \frac{\sigma\delta s_i}{(\sigma-\delta s_i-1)(\sigma-\delta s_i)}\hat{s}_i + \frac{\delta s_i(\sigma-2\delta s_i)}{(\sigma-\delta s_i-1)\sigma+\delta^2 s_i^2}\hat{s}_i.$$

This equation implies that the right-hand side of (1.22) is declining in r_i because Γ_i is declining in s_i and s_i is rising in r_i (see (1.11)). The former is seen from this equation by observing that $\sigma\delta s_i > \delta s_i(\sigma-2\delta s_i)$ and $(\sigma-\delta s_i-1)(\sigma-\delta s_i) < (\sigma-\delta s_i-1)\sigma+s_i^2\delta^2$. Collecting terms we can rewrite this equation as:

$$\hat{\Gamma}_i = -(\sigma-1)\,\hat{a}_i + \delta\hat{P} - \delta^2 s_i^2\frac{2(\sigma-\delta s_i-1)(\sigma-\delta s_i)+\sigma(\sigma-1)}{(\sigma-\delta s_i-1)(\sigma-\delta s_i)\left[(\sigma-\delta s_i-1)\sigma+\delta^2 s_i^2\right]}\hat{s}_i. \tag{A.1}$$

Next consider the total effect of a shift in the marginal cost a_i on Γ_i. From (1.11) we have:

$$\hat{s}_i = -\frac{\sigma-1}{1+(\sigma-1)\beta_i}\hat{a}_i = -\frac{(\sigma-1)(\sigma-\delta s_i-1)(\sigma-\delta s_i)}{(\sigma-\delta s_i-1)(\sigma-\delta s_i)+(\sigma-1)\delta s_i}\hat{a}_i.$$

Table A.1: *Average Number of Product Lines vs. Productivity Deciles*

Decile	*Log(Prod)*	*MeanInd*	*MeanSegs*
1	10.05	1.89	2.93
2	11.54	2.14	3.65
3	12.04	2.27	4.00
4	12.31	2.48	4.47
5	12.54	2.64	4.84
6	12.77	2.67	4.98
7	13.06	2.63	4.83
8	13.42	2.53	4.79
9	13.91	2.29	4.57
10	15.31	1.92	3.99

Note: This table shows the deciles of average log labor productivity for firms in the Compustat database for the year 2018, available through WRDS. Labor productivity is defined as the ratio of total sales to employment. It also shows the mean number of industries and business segments that are reported in the Compustat Segments Data. The data was accessed on June 2, 2020.

Substituting this expression into (A.1) we obtain the total impact of a_i on Γ_i:

$$\begin{aligned}\frac{\hat{\Gamma}_i}{(\sigma-1)\,\hat{a}_i} &= -1+\delta^2 s_i^2 \frac{2\,(\sigma-\delta s_i-1)\,(\sigma-\delta s_i)+\sigma\,(\sigma-1)}{\left[(\sigma-\delta s_i-1)\,\sigma+s_i^2\delta^2\right]^2}\\ &= \frac{(\sigma-1)\,s_i^2\delta^2-(\sigma-\delta s_i-1)^2\left(\sigma^2-\delta^2 s_i^2\right)}{\left[(\sigma-\delta s_i-1)\,\sigma+s_i^2\delta^2\right]^2}.\end{aligned}$$

It follows that a decline in the marginal cost a_i shifts upward the $\dot{\lambda}_i$=0 curve if and only if $(\sigma-1)\,s_i^2\delta^2 < (\sigma-\delta s_i-1)^2\left(\sigma^2-\delta^2 s_i^2\right)$.

A.2 Empirical Analysis

We now provide additional information on the empirical analysis. Table A.1 presents the data that has has been used to construct Figure 1.4 while Table A.2 presents the regression results. As pointed out in the main text, the coefficient for log productivity is positive and significantly different from zero and the coefficient for the square of log productivity is negative and significantly different from zero in both specifications; i.e., when we use the number of industries or the number of segments to measure a firm's product span. While in the main text we reported in Figure 1.3 the curvature of this quadratic form for the number

Table A.2: *Quadratic Relationship of Productivity on Product Span*

	Industries	Segments
log(Prod)	2.85**	5.50**
	(1.33)	(2.54)
$\log(\text{Prod})^2$	-0.11*	-0.21**
	(0.06)	(.11)
Primary Ind. FE	YES	YES
Obs	4126	4126
R^2	0.7334	0.4603

Robust standard errors clustered at the primary industry in parentheses.

*$p < 0.10$,**$p < 0.05$.

Note: This table shows the results of an OLS quadratic regression of the number of industries or segments on the log of labor productivity. The data includes all firms with positive sales and employment in the Compustat database for the year 2018. Labor productivity is defined as the ratio of total sales to employment. Segments here refers to the total number of business segments listed in the Compustat Segments Data by firm. The number of industries is the number of primary and secondary SIC codes listed across all business segments. We also include fixed effects for 4 digit primary SIC code listed on Compustat. Data was accessed on June 2, 2020.

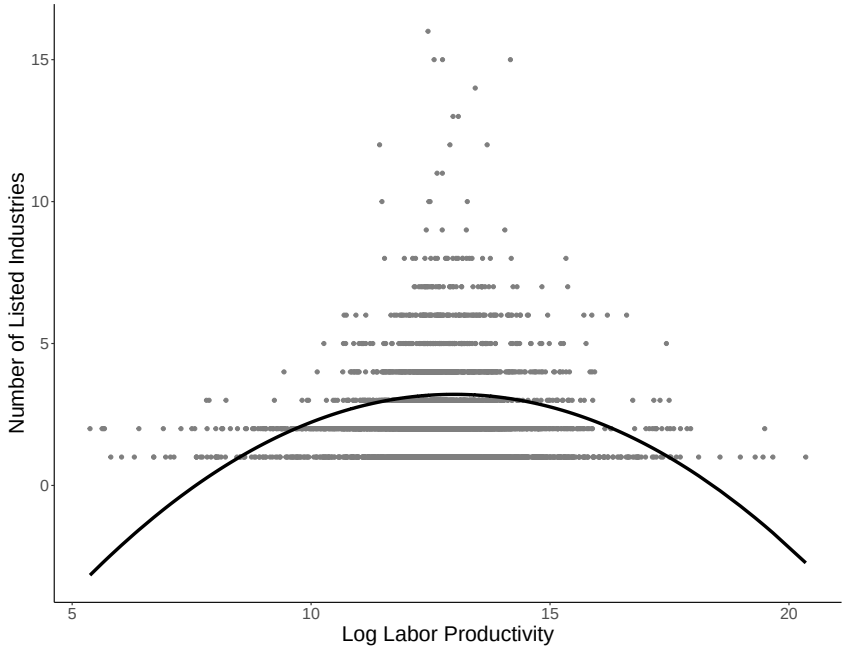


Figure A.1: *Number of Industries vs. Labor Productivity*

of segments as a proxy for the number of product lines, we now report a similar figure, Figure A.1, for the case in which the number of industries is used as a proxy for the number of product lines. As is evident, the two figures are quite similar.

A.3 Optimal Allocation

In Section 1.6 we characterize the optimal allocation, showing that it differs from the market outcome. In this part of the appendix we propose policies that implement the optimal allocation in a market economy with taxes and subsidies. In particular, we show that there exist consumer subsidies for the purchase of varieties of the differentiated product and corporate taxes on operating profits that lead to a market allocation that coincides with the optimal allocation. These taxes and subsidies are firm specific and they vary over time. Moreover, implementation of the optimal allocation requires the policy maker to commit to the entire time path of these taxes and subsidies, which vary across firms and across time.

Let $\overline{\gamma}$ be the factor that converts a producer price $\overline{p}$ into a consumer price $\overline{\gamma}\,\overline{p}$ and by γ_i the factor that converts a producer price p_i into a consumer price $\gamma_i p_i$. We allow these conversion factors to vary over time, although we will find that the optimal value of $\overline{\gamma}$ is constant. Importantly, both consumers and producers treat these factors as exogenous variables. A γ smaller than one represents a subsidy to consumers while a γ larger than one represents a tax. Finally, we denote by τ_i the factor that converts gross operating profits of firm i, $r_i P^{\delta} \left(\gamma_i p_i\right)^{-\sigma} \left(p_i - a_i\right)$, into net operating profits $\tau_i r_i P^{\delta} \left(\gamma_i p_i\right)^{-\sigma} \left(p_i - a_i\right)$. The factors τ_i may also vary over time, but the firms treat them as exogenous variables. A τ_i smaller than one represents a corporate tax on operating profits while a τ_i larger than one represents a corporate subsidy to operating profits.

With these policies in place, the demand for varies of the differentiated product (1.3) can be expressed as:

$$\overline{x} = P^{*\delta} \left(\overline{\gamma}\,\overline{p}\right)^{-\sigma},$$

$$x_i = P^{*\delta} \left(\gamma_i p_i\right)^{-\sigma},$$

where

$$P^* = \left[\overline{r} \left(\overline{\gamma}\, \overline{p}\right)^{1-\sigma} + \sum_{j=1}^{m} r_j \left(\gamma_i p_j\right)^{1-\sigma}\right]^{\frac{1}{1-\sigma}}.$$

In this exposition we use asterisks to denote equilibrium values of endogenous variables in the economy with taxes and subsidies. Large firms now maximize net operating profits $\tau_i r_i P^{*\delta} \left(\gamma_i p_i\right)^{-\sigma} \left(p_i - a_i\right)$ while small firms maximize operating profits $P^{*\delta} \left(\overline{\gamma}\, \overline{p}\right)^{-\sigma} \left(\overline{p} - \bar{a}\right)$. This yields the optimal pricing equations:

$$\overline{p}^* = \frac{\sigma}{\sigma - 1} \bar{a},$$

$$p_i^* = \frac{\sigma - \delta s_i^*}{\sigma - \delta s_i^* - 1} a_i, \tag{A.2}$$

where s_i^* is the share of consumer spending on goods of firm i, equal to

$$s_i^* = \frac{r_i \left(\gamma_i p_i^*\right)^{1-\sigma}}{P^{*1-\sigma}}. \tag{A.3}$$

We now propose the following numerical values of these policies:

$$\overline{\gamma} = \frac{\sigma - 1}{\sigma} \quad \text{and} \quad \gamma_i = \frac{\sigma - \delta s_i^* - 1}{\sigma - \delta s_i^*}, \tag{A.4}$$

which yields $\overline{\gamma}\, \overline{p} = \bar{a}$ and $\gamma_i p_j = a_j$. In other words, these policies lead to consumer prices that equal marginal costs of production. Note that every γ is smaller than one. Therefore consumers enjoy subsidies on all varieties of the differentiated product and the subsidies are larger on products with larger market shares.

With these subsidies a small firm's operating profits are $P^{*\delta} \left(\overline{\gamma}\, \overline{p}^*\right)^{-\sigma} \left(\overline{p}^* - \bar{a}\right)$, and free entry ensures that these profits equal the entry cost f. Using the firm's optimal pricing equation (A.2) and the subsidy policy (A.4), this free entry condition yields

$$\frac{1}{\sigma-1}P^{*\delta}\bar{a}^{1-\sigma}=f.$$

Comparing this to (1.33), we conclude that $P^*=C^*$, i.e., the price index equals the optimal resource cost of producing a unit of real consumption X. As a result, real consumption X is also at the optimal level, equal to $X^*=C^{*-\varepsilon}$, and the consumption levels of individual varies are at the optimal levels (see (1.34) and (1.35)):

$$\overline{x}^*=C^{*\delta}\overline{a}^{-\sigma}=(\sigma-1)\overline{a}^{-1}f,$$

$$x_i^*=C^{*\delta}a_i^{-\sigma}=(\sigma-1)\overline{a}^{\sigma-1}a_i^{-\sigma}f.$$

It remains to examine the investment policies of large firms.

Recognizing that $P^*=C^*$ is constant on the dynamic path, (A.2) and (A.3) implicitly define the optimal price of firm i as a function of its product span, $p_i^*(r_i)$, similarly to the analysis of the market economy without government intervention. The only difference is that now there are policy instruments that the firms treat as exogenous. As a result, profits of firm i net of taxes and investment costs are

$$\pi_i(\iota_i,r_i)=\tau_i r_i C^{*\delta}\left[\gamma_i p_i^*(r_i)\right]^{-\sigma}\left[p_i^*(r_i)-a_i\right]-\iota_i$$

and the current value Hamiltonian of the firm's optimal control problem is

$$\mathcal{H}(\iota_i,r_i,\lambda_i)=\tau_i r_i C^{*\delta}\left[\gamma_i p_i^*(r_i)\right]^{-\sigma}\left[p_i^*(r_i)-a_i\right]-\iota_i+\lambda_i\left[\phi(\iota_i)-\theta r_i\right].$$

The first-order conditions for the optimal control problem are therefore:

$$\frac{\partial\mathcal{H}}{\partial\iota_i}=-1+\lambda_i\phi'(\iota_i)=0,$$

$$-\frac{\partial\mathcal{H}}{\partial r_i}=-\tau_i\Gamma_i^*(r_i)+\theta\lambda_i=\dot{\lambda}_i-\rho\lambda_i,$$

where

$$\Gamma_i^*(r_i) \equiv \frac{\partial \left\{ r_i C^{*\delta} \left[\gamma_i p_i^*(r_i) \right]^{-\sigma} \left[p_i^*(r_i) - a_i \right] \right\}}{\partial r_i},$$

and the transversality conditions are:

$$\lim_{t \to \infty} e^{-\rho t} \lambda_i(t) r_i(t) = 0.$$

Now recall that the optimal investment in innovation is constant on the dynamic path and satisfies $\lambda_i^* \phi'(\iota_i^*) = 1$, where λ_i^* is is given in (1.39), i.e.,

$$\lambda_i^* = \frac{1}{\rho + \theta} \left(\frac{\overline{a}}{a_i} \right)^{\sigma - 1} f.$$

The first-order conditions of the firm's optimal control problem imply that this investment pattern is attained if and only if:

$$\tau_i \Gamma_i^*(r_i^*) = \left(\frac{\overline{a}}{a_i} \right)^{\sigma - 1} f \tag{A.5}$$

at every point in time, where r_i^* is the optimal product span. It follows from this result that operating profits of firm i are taxed ($\tau_i < 1$) if and only if

$$\Gamma_i^*(r_i^*) > \left(\frac{\overline{a}}{a_i} \right)^{\sigma - 1} f.$$

We show next that $\tau_i \epsilon (0,1)$; that is, the optimal policy consists of taxing operating profits.

First note that

$$\frac{\Gamma_i^*(r_i)}{C^{*\delta}} = \frac{\partial \left\{ r_i \left[\gamma_i p_i^*(r_i) \right]^{-\sigma} \left[p_i^*(r_i) - a_i \right] \right\}}{\partial r_i}$$

$$= \left[\gamma_i p_i^*(r_i) \right]^{-\sigma} \left[p_i^*(r_i) - a_i \right] - \left\{ \sigma r_i \gamma_i^{-\sigma} p_i^*(r_i)^{-\sigma - 1} \left[p_i^*(r_i) - a_i \right] - r_i \left[\gamma_i p_i^*(r_i) \right]^{-\sigma} \right\} \frac{\partial p_i^*(r_i)}{\partial r_i}$$

$$= \gamma_i^{-\sigma} p_i^*(r_i)^{-\sigma} \left(\left[p_i^*(r_i) - a_i \right] - \left\{ \sigma r_i p_i^*(r_i)^{-1} \left[p_i^*(r_i) - a_i \right] - r_i \right\} \frac{\partial p_i^*(r_i)}{\partial r_i} \right)$$

$$= \gamma_i^{-\sigma} p_i^* (r_i)^{-\sigma} \left[\frac{1}{\sigma - \delta s_i^* (r_i) - 1} a_i - r_i \left[\frac{\sigma}{\sigma - \delta s_i^* (r_i)} - 1 \right] \frac{\partial p_i^* (r_i)}{\partial r_i} \right]$$

$$= \gamma_i^{-\sigma} p_i^* (r_i)^{-\sigma} \left[\frac{1}{\sigma - \delta s_i^* (r_i) - 1} a_i - r_i \frac{\delta s_i^* (r_i)}{\sigma - \delta s_i^* (r_i)} \frac{\partial p_i^* (r_i)}{\partial r_i} \right].$$

However,

$$\frac{\partial p_i^* (r_i)}{\partial r_i} \frac{r_i}{p_i^*} = \frac{\beta_i^* (r_i)}{1 + (\sigma - 1)\beta_i^* (r_i)}$$

where

$$\beta_i^* (r_i) = \frac{\delta s_i^* (r_i)}{\left[\sigma - \delta s_i^* (r_i) - 1\right] \left[\sigma - \delta s_i^* (r_i)\right]}.$$

Therefore

$$\begin{aligned} \Gamma_i^* (r_i) &= C^{*\delta} \gamma_i^{-\sigma} p_i^* (r_i)^{-\sigma} \left[\frac{1}{\sigma - \delta s_i^* (r_i) - 1} a_i - \left[p_i^* (r_i) \frac{\delta s_i^* (r_i)}{\sigma - \delta s_i^* (r_i)} \right] \frac{\beta_i^* (r_i)}{1 + (\sigma - 1)\beta_i^* (r_i)} \right] \\ &= \gamma_i^{-\sigma} a_i^{1-\sigma} C^{*\delta} \left[\frac{\sigma - \delta s_i^* (r_i)}{\sigma - \delta s_i^* (r_i) - 1} \right]^{-\sigma} \frac{\sigma}{\left[\sigma - \delta s_i^* (r_i) - 1\right] \sigma + \delta^2 s_i^* (r_i)^2} \\ &= \gamma_i^{-\sigma} \left(\frac{\overline{a}}{a_i} \right)^{\sigma - 1} f \left[\frac{\sigma - \delta s_i^* (r_i)}{\sigma - \delta s_i^* (r_i) - 1} \right]^{-\sigma} \frac{\sigma (\sigma - 1)}{\left[\sigma - \delta s_i^* (r_i) - 1\right] \sigma + \delta^2 s_i^* (r_i)^2}, \end{aligned}$$

where we used (1.33) in deriving the last line. Now compare this formula to (1.20). Since we showed that the expression on the right-hand side of (1.20) declines in r_i, it follows that—holding γ_i constant—$\Gamma_i^* (r_i)$ also declines in r_i. This ensures concavity in r_i of the firm's decision problem.

Finally, we show that $\tau_i \epsilon (0, 1)$ in every time period, implying that the optimal policy consists of a tax on operating profits. To this end use the formula for the subsidy factor γ_i

together with the optimal tax formula (A.5) to obtain:

$$\tau_i = \Gamma_i^*(r_i)^{-1}\left(\frac{\overline{a}}{a_i}\right)^{\sigma-1} f = \frac{\sigma\left[\sigma - \delta s_i^*(r_i) - 1\right] + \delta^2 s_i^*(r_i)^2}{\sigma(\sigma-1)},$$
$$= 1 - \frac{\left[\sigma - \delta s_i^*(r_i)\right]\delta s_i^*(r_i)}{\sigma(\sigma-1)},$$

which shows that $\tau_i \epsilon (0,1)$ at every point in time.

For a firm with rising product span the share of consumer spending on its products rises over time, i.e., $s_i^*(r_i)$ is an increasing function. Therefore the corporate tax rate is rising over time (τ_i is decreasing) if and only $\sigma > 2\delta s_i^*(r_i)$. Since $\delta = \sigma - \varepsilon > 0$, it follows that for $\sigma > 2\varepsilon$ there exists a market share $s_c = \sigma/2(\sigma - \varepsilon)$ such that the tax rate is rising for market shares below s_c and declining for larger market shares. In the opposite case, when $\sigma > 2\varepsilon$, the corporate tax rate always rises for firms that expand their product span.

A.4 Comparative Statics: Given Number of Brands

In this section we examine the case in which the number of single-product firms, $\overline{r}$, as well the number of products available to each one of the large firms, r_i, are given. Equations (1.7) and (1.8) imply:

$$\hat{p}_i = \hat{a}_i + \frac{\delta s_i}{(\sigma - \delta s_i - 1)(\sigma - \delta s_i)}\hat{s}_i, \tag{A.6}$$

$$\hat{s}_i = \hat{r}_i - \sum_{j=1}^{m} s_j \hat{r}_j - (\sigma - 1)(\hat{p}_i - \sum_{j=1}^{m} s_j \hat{p}_j).$$

Substituting the last equation into (A.6) yields:

$$[1 + \beta_i(\sigma - 1)]\hat{p}_i - \beta_i(\sigma - 1)\sum_{j=1}^{m} s_j \hat{p}_j = \hat{a}_i + \beta_i(\hat{r}_i - \sum_{j=1}^{m} s_j \hat{r}_j), \text{ for all } i.$$

These equations can also be expressed as:

$$\mathbf{B}\hat{\mathbf{p}} = \mathbf{R}\hat{\mathbf{r}} + \hat{\mathbf{a}}, \tag{A.7}$$

where $\mathbf{B}$ is an $m \times m$ matrix with elements:

$$b_{ii} = 1 + \beta_i(\sigma - 1)(1 - s_i),$$

$$b_{ij} = -\beta_i(\sigma - 1)s_j, \text{ for } j \neq i,$$

$\hat{\mathbf{p}}$ is an $m \times 1$ column vector with elements p_i, where a hat represents a proportional rate of change (i.e., $\hat{p}_i = dp_i/p_i$), $\mathbf{R}$ is an $m \times m$ matrix with elements:

$$r_{ii} = \beta_i(1 - s_i),$$

$$r_{ij} = -\beta_i s_j, \text{ for } j \neq i,$$

$\hat{\mathbf{r}}$ is an $m \times 1$ column vector with elements $\hat{r}_i$, where a hat represents a proportional rate of change, and $\hat{\mathbf{a}}$ is an $m \times 1$ column vector with elements $\hat{a}_i$, where a hat represents a proportional rate of change.

Since

$$|b_{ii}| - \sum_{j \neq i} |b_{ij}| = 1 + \beta_i(\sigma - 1)(1 - \sum_{j=1}^{m} s_j) > 1,$$

$\mathbf{B}$ is a diagonally dominant matrix with positive diagonal and negative off-diagonal elements. It therefore is an M-matrix and its inverse has all positive entries. This inverse, denoted by $\tilde{\mathbf{B}} = \mathbf{B}^{-1}$, is therefore an $m \times m$ matrix with elements $\tilde{b}_{ij} > 0$. Next note that $\mathbf{B}$ can be expressed as:

$$\mathbf{B} = \mathbf{I} + (\sigma - 1)\mathbf{R},$$

where $\mathbf{I}$ is the identity matrix. Therefore:

$$\mathbf{B}^{-1}\mathbf{B} = \tilde{\mathbf{B}} + (\sigma - 1)\tilde{\mathbf{B}}\mathbf{R} = \mathbf{I}. \tag{A.8}$$

It follows from this equation that:

$$\tilde{b}_{ii} + (\sigma - 1) \sum_{j=1}^{m} \tilde{b}_{ij} r_{ji} = 1,$$

$$\tilde{b}_{ik} + (\sigma - 1) \sum_{j=1}^{m} \tilde{b}_{ij} r_{jk} = 0, \text{ for } k \neq i.$$

Summing these up yields:

$$\sum_{k=1}^{m} \tilde{b}_{ik} + (\sigma - 1) \sum_{j=1}^{m} \tilde{b}_{ij} \sum_{k=1}^{m} r_{jk} = 1, \text{ for all } i. \tag{A.9}$$

Since:

$$\sum_{k=1}^{m} r_{jk} = \beta_j (1 - \sum_{k=1}^{m} s_k) > 0$$

and $\tilde{b}_{ik} > 0$ for all i and k, it follows from (A.9) that:

$$0 < \tilde{b}_{ik} < 1 \text{ for all } i \text{ and } k.$$

Equation (A.8) implies:

$$(\sigma - 1)\tilde{\mathbf{B}}\mathbf{R} = \mathbf{I} - \tilde{\mathbf{B}},$$

and therefore $\tilde{\mathbf{B}}\mathbf{R}$ has positive diagonal elements and negative off-diagonal elements.

Going back to the comparative statics equations (A.7), we have:

$$\hat{\mathbf{p}} = \tilde{\mathbf{B}}\mathbf{R}\hat{\mathbf{r}} + \tilde{\mathbf{B}}\hat{\mathbf{a}}.$$

It follows from the properties of $\tilde{\mathbf{B}}$ that a decline in a_i reduces every price p_j, but less than proportionately. Equation (A.6) than implies that all market share $s_j, j \neq i$, decline while the market share s_i rises. And it follows from the properties of $\tilde{\mathbf{B}}\mathbf{R}$ and (A.6) that an increase in r_i raises the price and market share of firm i and reduces the price and market share of every other firm $j \neq i$. Noting that the markup of every firm i is larger the larger its market

share, we therefore have:

Proposition 11. *Suppose that the number of firms and their product ranges are given. Then: (i) an increase in r_i raises the price, markup and market share of firm i, and reduces the price, markup and market share of every other large firm; (ii) a decline in a_i reduces the price of every large firm less than proportionately, raises the markup and market share of firm i, and reduces the markup and market share of every other large firms.*

A.4.1 Aggregative Economy

In this section we show how to construct an aggregative economy with a continuum of industries, each one of the type analyzed in the main text of this paper.

We consider an economy with a continuum of individuals of mass 1, each one providing one unit of labor. The labor market is competitive and every individual earns the same wage rate.

There is a continuum of sectors of measure one, each one producing a differentiated product. Real consumption in sector k is:

$$X^k = \left[\int_0^{N^k} x^k(\omega)^{\frac{\sigma^k - 1}{\sigma^k}} d\omega \right]^{\frac{\sigma^k}{\sigma^k - 1}}, \ \sigma^k > 1,$$

where N^k is the number of varieties available in sector k, $x^k(\omega)$ is consumption of variety ω in sector k, and σ^k is the elasticity of substitution in sector k. Using this definition, the price index of X^k is:

$$P^k = \left[\int_0^{N^k} p^k(\omega)^{1-\sigma^k} d\omega \right]^{\frac{1}{\sigma^k - 1}},$$

where $p^k(\omega)$ is the price of variety ω. The log utility of a representative individual is:

$$log\,(u) = \int_0^1 log\left(X^k\right) dk.$$

In these circumstances every individual spends an equal amount of money in every sector.

Therefore, if E denotes aggregate spending per capita, spending per capita in sector k also equals E. In this event, aggregate demand for variety ω in sector k is:

$$x^k(\omega) = A^k p(\omega)^{-\sigma}, \tag{A.10}$$

$$A^k = E\left(P^k\right)^{\sigma^k - 1}. \tag{A.11}$$

An individual's inter-temporal utility function is:

$$U = \int_0^\infty e^{-\rho t} log\left(u_t\right) dt,$$

where ρ is the subjective discount rate. As a result, the intertemporal allocation of spending satisfies:

$$\frac{\dot{E}_t}{E_t} = \zeta_t - \rho, \tag{A.12}$$

where ζ_t is the interest rate at time t.

Two types of firms operate in sector k: atomless single-product firms and large multi-product firms, each one with a positive measure of product lines. Single-product firms produce $\bar{r}^k > 0$ varieties, each one specializing in a single brand. Large firm i in sector k has $r_i^k > 0$ product lines, $i = 1, 2, ..., m^k$, where m^k is the number of large firms in this sector. All the brands supplied to the market are distinct from each other.

All single-product firms share the same technology, which requires $\bar{a}^k$ unit of labor per unit output in sector k. Facing the demand function (A.10), a single-product firm maximizes profits $A^k p(\omega)^{-\sigma}\left[p(\omega) - \bar{a}^k\right]$, taking as given the demand shifter A^k. Therefore, a single-product firm prices its brand ω according to $p(\omega) = \overline{p}^k$, where:

$$\overline{p}^k = \frac{\sigma^k}{\sigma^k - 1}\bar{a}^k. \tag{A.13}$$

This yields the standard markup $\overline{\mu}^k = \sigma^k / \left(\sigma^k - 1\right)$ for a monopolistically competitive firm.

A large firm i has a technology that requires a_i^k units of labor per unit output, and it faces the demand function (A.10) for each one of its brands. As a result, it prices every brand equally. We denote this price by p_i^k. The firm chooses p_i^k to maximize profits

$r_i^k A^k p_i^{-\sigma} \left(p_i - a_i^k\right)$. However, unlike a single-product firm, a large firm does not view A^k as given, because it recognizes that

$$P^k = \left(\bar{r}^k \left(\bar{p}^k\right)^{1-\sigma} + \sum_{j=1}^{m^k} r_j^k \left(p_j^k\right)^{1-\sigma^k}\right)^{\frac{1}{1-\sigma^k}}, \tag{A.14}$$

and therefore that its pricing policy has a measurable impact on the price index of the differentiated product. It takes, however, the spending level E as given, because sector k is of measure zero. Accounting for this dependence of P^k on the firm's price, the profit maximizing price is:

$$p_i^k = \frac{\sigma^k - \left(\sigma^k - 1\right) s_i^k}{\left(\sigma^k - 1\right)\left(1 - s_i^k\right)} a_i^k, \tag{A.15}$$

where s_i^k is the market share of firm i in sector k and:

$$s_i^k = \frac{r_i^k \left(p_i^k\right)^{1-\sigma^k}}{\left(P^k\right)^{1-\sigma^k}} = \frac{r_i^k \left(p_i^k\right)^{1-\sigma^k}}{\bar{r}^k \left(\bar{p}^k\right)^{1-\sigma} + \sum_{j=1}^{m^k} r_j^k \left(p_j^k\right)^{1-\sigma^k}}. \tag{A.16}$$

Equations (A.15) and (A.16) jointly determine prices and market shares of large firms. The markup factor of firm i is $\mu_i^k = \left[\sigma^k - \left(\sigma^k - 1\right) s_i^k\right] / \left[\left(\sigma^k - 1\right)\left(1 - s_i^k\right)\right]$, which is increasing in its market share. When the market share equals zero the markup is $\sigma^k/(\sigma^k - 1)$, the same as the markup of a single product firm. The markup factor varies across firms as a result of differences in either the product span, r_i^k, or the marginal production cost, a_i^k. We analyze the dependence of prices, market shares and markups on marginal costs and product spans in the next section.

Entry of Single-Product Firms

The number of large firms in every sector, m^k, is given. Unlike large firms, however, single-product firms enter the industry until their profits equal zero. In every sector the firms play a two-stage game: in the first stage single-product firms enter; in the second stage all firms play a Bertrand game as described above. Under these circumstances, (A.13) and (A.15) portray the equilibrium prices, except that the number of single product firms, $\bar{r}^k$, is

endogenous. We seek to characterize a subgame perfect equilibrium of this game.

To determine the equilibrium number of single-product firms, assume that they face an entry cost f^k in sector k and they enter until profits equal zero. In a subgame perfect equilibrium every entrant correctly forecasts aggregate spending on the sector's products, the number of entrants, and the price that will be charged for every variety in the second stage of the game. Therefore, every single-product firm correctly forecasts the price index and A^k. Using the optimal price (A.13) and the profit function $A^k p(\omega)^{-\sigma}\left[p(\omega)-\bar{a}^k\right]$, this free entry condition can be expressed as:

$$\frac{1}{\sigma^k}A^k\left(\frac{\sigma^k}{\sigma^k-1}\bar{a}^k\right)^{1-\sigma^k}=f^k. \tag{A.17}$$

The left-hand side of this equation describes the operating profits, which equal a fraction $1/\sigma^k$ of revenue, while the right-hand side represents the entry cost. In these circumstances the demand shifter A^k is determined by f^k and $\bar{a}^k$, and it is rising in both f^k and $\bar{a}^k$. Importantly, it does not depend on the number of large firms nor on their product spans. Moreover, given the spending level E, which is determined at the economy-wide level and is not influenced by product spans in sector k (because the sector is of measure zero), the price index P^k is also independent of product spans in sector k. In particular, changes over time in this price index are driven by changes in aggregate spending. For this reason (A.11) and (A.12) imply:

$$\frac{\dot{P}_t^k}{P_t^k}=\frac{1}{\sigma^k-1}\left(\rho-\zeta_t\right). \tag{A.18}$$

Optimal Control

We can now compute the response of p_i^k and s_i^k to changes in r_i^k as we did in the main text, and use the solution in the firm's optimal control problem. In the optimal control problem large firm i in sector k takes as given the path of the interest rate r_t and the path of spending E_t. After characterizing this solution we can use it to express the market clearing conditions. Spending E_t has to equal wage income and aggregate profits net of investment costs. This will give us the growth model. If we use the formulation from the main text, the steady state

will have zero growth. But one could add a long-run growth mechanism, such as declining costs of innovation as a function of the cumulative experience in innovation, as is Romer (1990). The steady state should be easy to analyze in either case.

As in the main text, investment is given by

$$\dot{r}_i^k = \phi(\iota_i^k) - \theta r_i^k, \text{ for all } t \geq 0, \tag{A.19}$$

At every point in time the firms play a two stage game. In the first stage single-product firms enter and large firms invest in innovation. Single-product firms live only one instant of time. For this reason they make profits only in this single instant. Under the circumstances the demand shifter A^k is determined by the free entry condition, and it remains constant as long as the cost of entry and the cost of production of the single-product firms do no change. It follows that the profit flow of large firm i is:

$$\pi_i^k = r_i^k A^k \left(p_i^k\right)^{-\sigma} \left(p_i^k - a_i^k\right) - \iota_i^k, \text{ for all } t \geq 0,$$

where A^k is the same at every t while π_i^k, r_i^k, p_i^k and ι_i^k change over time, and p_i^k is given by $p_i^k = \frac{\sigma^k - (\sigma^k - 1)s_i^k}{(\sigma^k - 1)\left(1 - s_i^k\right)} a_i^k$. We can write the optimal control problem as:

$$\max_{\left\{\iota_i^k(t), r_i^k(t)\right\}_{t \geq 0}} \int_0^\infty e^{-\int_0^t \zeta_\tau d\tau} \pi_i^k \left[\iota_i^k(t), r_i^k(t)\right] dt$$

The main difference between this formulation and the formulation in the main text is that now we no longer have $\zeta_t = \rho$ at each point in time, but rather $\zeta_t = \frac{\dot{E}_t}{E_t} + \rho$. The current-value Hamiltonian of this problem is:

$$\mathcal{H}(\iota_i^k, r_i^k, \lambda_i^k) = \left\{r_i^k A^k p_i^k (r_i)^{-\sigma} \left[p_i^k (r_i) - a_i^k\right] - \iota_i^k\right\} + \lambda_i^k \left[\phi\left(\iota_i^k\right) - \theta r_i^k\right],$$

and the first-order conditions are:

$$\frac{\partial \mathcal{H}}{\partial \iota_i^k} = -1 + \lambda_i^k \phi' \left(\iota_i^k\right) = 0,$$

$$-\frac{\partial \mathcal{H}}{\partial r_i^k} = -\frac{\partial \left[r_i^k A^k \left(p_i^k\right)^{-\sigma} \left(p_i^k - a_i^k\right)\right]}{\partial r_i^k} + \theta \lambda_i^k = \dot{\lambda}_i^k - \zeta_t \lambda_i^k.$$

Note that the path of the price index P_t^k is determined by the growth rate of the aggregate economy that each firm takes as exogenous. Therefore, the resulting first-order conditions have a similar form to those we derived in the main text:

$$\lambda_i^k \phi' \left(\iota_i^k\right) = 1, \tag{A.20}$$

$$\dot{\lambda}_i^k = \left(\zeta_t + \theta\right) \lambda_i^k - A^k p_i^k \left(r_i^k\right)^{-\sigma^k} \left\{ p_i \left(r_i^k\right) - a_i^k - r_i^k \left(\sigma^k p_i^k \left(r_i^k\right)^{-1} \left[p_i^k \left(r_i^k\right) - a_i^k\right] - 1\right) \frac{dp^k \left(r_i^k\right)}{dr_i^k} \right\}. \tag{A.21}$$

Substituting (A.20) into (A.19) yields:

$$\dot{r}_i = \phi \left[\iota_i \left(\lambda_i\right)\right] - \theta r_i. \tag{A.22}$$

The second differential equation is obtained by substituting the pricing equation into (A.21):

$$\dot{\lambda}_i^k = \left(\zeta_t + \theta\right) \lambda_i^k - \Gamma_i^k \left(r_i^k\right), \tag{A.23}$$

where:

$$\Gamma_i^k \left(r_i^k\right) \equiv a_i^{1-\sigma^k} A^k \sigma \left[\frac{\sigma^k - (\sigma^k - 1) s_i^k \left(r_i^k\right)}{\left(\sigma^k - 1\right)\left(1 - s_i^k\right)}\right]^{-\sigma^k} \frac{1}{\left(\sigma^k - 1\right)\left(1 - s_i^k\right)\sigma + s_i \left(r_i\right)^2 (\sigma - 1)^2}. \tag{A.24}$$

Thus, our two differential equations are similar to the main text, with the caveat that

the interest rate is evolving over time. Specifically, the dynamics are such that aggregate spending must satisfy $\zeta_t = \frac{\dot{E}_t}{E_t} + \rho$.

In steady state:

$$\phi\left[\iota_i^k\left(\lambda_i^k\right)\right] = \theta r_i^k, \tag{A.25}$$

$$(\rho + \theta)\,\lambda_i^k = \Gamma_i^k\left(r_i^k\right), \tag{A.26}$$

where we have used the fact that in steady state $\zeta_t = \rho$. The comparative statics of this system have the same form as in the main text. But note that while the key condition for having an inverted-U relationship between productivity and product span was $(\sigma - \delta - 1)^2\left(\sigma^2 - \delta^2\right) < (\sigma - 1)\,\delta^2$ in the main text, the formula is the same now with the exception that δ is replaced with $\sigma - 1$. This reduces the condition to $0 < \left(\sigma^k - 1\right)^3$, which is always satisfied. Thus, in this formulation we would expect every sector to have the inverted-U property. Another comparative static to note is the effect of an increase in the steady state expenditure level E. This shifts upward the curve associated with (A.26) in the phase diagram, resulting in an instantaneous increase in λ_i^k and a trajectory of further expansion of r_i^k and rising profits. Thus, firms growing in other sectors reinforce the market dominance of large firms across industries through a pecuniary externality.

In order to close the model we need to solve for the steady state expenditure level. The market clearing condition is simply that revenue must equal net profits plus the total wage bill. With a unit mass of labor and the wage rate as the numeraire, the resulting condition takes the form:

$$E_t = 1 + \int_{k\in K}\left[\sum_{i=1}^{m^k} r_i^k A^k\left(p_i^k\right)^{-\sigma}\left(p_i^k - a_i^k\right) - \iota_i^k\right] dk. \tag{A.27}$$

We can further simplify this by recalling that $A^k = E_t\left(P^k\right)^{\sigma^k - 1}$. This means that we can use (A.27) to obtain:

$$E_t = \left[1 - \int_{k \in K} \left[\sum_{i=1}^{m^k} r_i^k \left(P^k\right)^{\sigma^k - 1} \left(p_i^k\right) \left(p_i^k - a_i^k\right) - \iota_i^k\right] dk\right]^{-1}. \tag{A.28}$$

Thus, the steady state expenditure level is increasing in the net profits of large firms across sectors. This equation also holds at every point in time, noting that the optimal investment levels depend on the path of aggregate expenditure through the interest rate. It follows that in order to solve the path of spending we need to ensure that the paths of profits of all firms aggregates to the path that rationalizes the optimal investments at each point in time.

Appendix B

Appendix to Chapter 2

B.1 Why Focus on Learning

B.1.1 Major Switching Between Freshman and Senior Year

One way to test the informational channel is to consider what students learn once they enter college. As shown in Crossley *et al.* (2021), students update their beliefs throughout college, developing more accurate predictions. Given the evidence that expectations of future earnings affect a student's choice of major, it is unsurprising that many students switch their majors after acquiring new information. Looking at the College Senior Survey performed by HERI which can be linked to the True Freshman Survey, I observe that roughly 40% of students switch majors. This is roughly in line with data from the National Center for Education Statistics. We might expect students with more information about wages for their stated major to be less likely to switch majors as they are less likely to experience a negative shock to their expectations.

In line with the hypothesis that local labor markets provide information about potential majors, table B.1 shows that students that state their intentions to major in a field which is more highly represented in their local labor market have a significantly lower likelihood of switching out of this degree. This is suggestive that students may be updating less about their intended major when it is well represented in their local labor market. The effect

Table B.1: *Effect of Labor Market Similarities on Major Switching*

	Fraction Switching		
Fraction of Workers with Major in County	-0.87***	-0.60***	-0.53**
	(0.09)	(0.18)	(0.18)
Major FE	NO	YES	YES
State FE	NO	NO	YES
Obs	68142	68142	68142
R^2	1.5%	4.2%	4.4%

Robust standard errors clustered at the primary industry in parentheses.

$^{*}p < 0.05$, $^{**}p < 0.01$,$^{***}p < 0.001$.

remains significant with a similar point estimate after controlling for major or state fixed effects.

B.1.2 Information Flows Through Local Major and Industry Composition

In an attempt to distinguish the underlying mechanism behind the informational signals from local labor markets, I compare the importance of the local educational, occupational and industrial composition on migration flows. Counties which are more similar in terms of these factors might be expected to have more migration between them. This may be due to companies relocating workers across cities or because skills learned on the job are more transferable across cities. However, it might also be the case that informational flows through these channels such that workers feel more confident migrating to areas with similar labor market structures. For instance, someone in Boston which has a relatively high density of software engineers is more likely to know the wage of software engineers in San Francisco. This is likely true across occupations, industries and degrees.

As a test to see whether this is true in the data I form a simple measure of proximity in terms of occupation, industry and education. Define matrix O, such that each column represents the occupation distribution in a particular county, c. Therefore O_{co} is the fraction of prime aged workers in county c with occupation o. The likelihood that someone in that occupation receives information about a different occupation is proxied by the occupational

transition matrix between the two which is estimated from the CPS. Thus, for occupation we define the proximity in terms of occupation between c and c' as

$$\Sigma^{o}_{cc'} = O_c T O_{c'}$$

Additionally, the total level of migration is going to depend on the population in c, L_c, and in c', $L_{c'}$, as well as the distance between locations. Using Census data for the year 2000, I estimate the following regression:

$$\log(M(c,c')) = \alpha + \beta \log(\Sigma) + \beta_2 \log(L_c) + \beta_3 \log(L_c)' + \eta \log(\delta(c,c'))$$

where $M(c,c')$ is the number of movers and $\delta(c,c')$ is the distance between c and c'. The results of this regression focusing on movers between 25 and 35 years old are shown in table B.2.

Table B.2: *Migration Based on Labor Market Similarities*

	Movers with $25 \leq$ Age ≤ 35				Age > 35
Occupational Similarity	0.36**			0.21·	0.14***
	(0.12)			(0.12)	(0.03)
Industrial Similarity		0.31***		0.25**	0.14***
		(0.07)		(0.08)	(0.02)
Educational Similarity			0.18**	0.13*	-0.00
			(0.06)	(0.06)	(0.01)
Obs	16583	16583	16583	16583	
R^2	1.5%	4.2%	4.4%		

Robust standard errors clustered at the primary industry in parentheses.

$^{*}p < 0.05$, $^{**}p < 0.01$,$^{***}p < 0.001$.

I focus on this age range as it is following graduation and is most likely to be a move

that could have been anticipated when making educational decisions. Column 4 shows that when taken together industrial and educational similarities are more important determinants of migration patterns for young movers. When I limit the sample to movers who are older than 35 years of age, the impact of educational similarity disappears whereas occupational and industrial composition becomes more significant. This is what you would predict if skills acquired overtime in a particular occupation or industry become more important determinants of the ability to find a new job elsewhere. However, the evidence for young workers where these skills haven't been built up are more likely to be due to information especially in light of the information above that suggests that along with developing skills in college, students are improving their estimate of expected earnings and that this extends to expected earnings across locations.

Appendix C

Appendix to Chapter 3

C.1 Numerical Example: The Uniform Commercial Code and "Reasonable" Defaults

In this section, we discuss the role of default rules in a slight modification to a canonical incomplete contracting problem, which closely follows an example in Easterbrook and Fischel (1982).

A buyer and a seller enter into a delivery contract for a widget. The contract specifies the time to delivery q as well as the price c. The seller invests a fixed cost of k to produce the good.

There is some uncertainty about an ex-ante indescribable but ex-post verifiable state of the world $\omega \in \{\omega_1, \omega_2\}$ which affects the time it takes the seller to deliver the widget. Suppose ω_1 is a "bad" verifiable state of the world and ω_2 is a "good" verifiable state of the world. For instance, a buyer and seller contracting in early 2021 may not have been able to foresee, and therefore contract on, a "bad" event ω_1 in which a 1300-ft container ship blocks the Suez Canal for six days.

Furthermore, we assume that there is an unverifiable component of the state of the world which affects the benefits from trade, represented by $\theta \in \{\theta_1, \theta_2\}$. The timing is as follows:

- At t_0, the buyer and seller sign an initial contract (q, c).

- At t_1, the seller invests k.
- At t_2, ω (ex-ante indescribable but ex-post verifiable) is revealed, θ (observable but unverifiable) is revealed.
- At t_3, the buyer and seller renegotiate to $h((q,c),k,\omega,\theta)$ where h is an arbitrary efficient bargaining function.

Contractual possibilities are limited by the fact that the buyer and seller cannot contract on ω or θ at t_0, and that, although they can renegotiate at t_3, they cannot enter the relationship at t_3 (because the seller must make a relationship-specific investment before ω and θ are realized). As in Aghion *et al.* (1994) the buyer and seller could agree to an initial contract (q_0, c_0) from which they can negotiate once ω and θ are revealed. This sequence is shown in Figure C.1a.

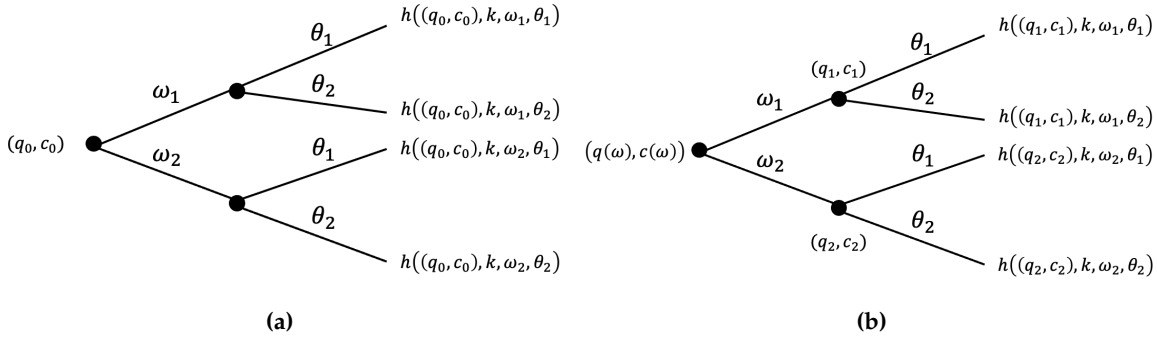


Figure C.1: *(a) Bargaining from an Explicit Prespecified Contract (q_0, c_0); (b) Bargaining from the U.C.C.'s "Reasonable" Default $(q_d(\omega), c_d(\omega))$*

The Uniform Commercial Code's default rules governing delivery times offer the buyer and the seller another option. The UCC states that "time for shipment or delivery... if not provided in this article or agreed upon shall be a reasonable time." If the buyer and seller choose not to specify a delivery time at t_0, they know that the default rule—will fill in the gap with a "reasonable" delivery time. Crucially, when the gap is filled in, at t_2, the regulator *will use all verifiable information* that is available at that time, including the realization of ω. So the contract that the parties implicitly sign when they leave out a delivery time is a contract with the default delivery time, which is function of ω. This case is shown in

Figure C.1b.

To understand the benefits of the U.C.C.'s default we look at a simple numerical example. We focus on the "bad" state of the world where $\omega = \omega_1$, and consider buyer and seller payoffs in t_3 when the initial contract is fully determined and when it is left open.

The buyer and seller's final t_3 payoffs have three components: 1) the payoff they would have gotten under the default contract, 2) the total surplus generated after efficient renegotiation, and 3) their investment costs. Consider the valuations for the buyer and seller under state θ_1 to be given by $V_0(\theta_1) = (2, -2)$ and $V_0(\theta_2) = (1, -1)$. The total surplus generated in state θ_1 is given by $S(\theta_1) = 4$ and $S(\theta_2) = 6$. Lastly, suppose that the seller has to pay an up front cost of 1 in every state so that the cost to buyer and seller are given by $C = (0, 1)$. Assume that the buyer and the seller have equal bargaining power, so that the gains from renegotiating the default contract will be split evenly.

When the buyer and seller write an explicit delivery contract (q_0, c_0), their payoffs are shown in Figure C.2a. In this case, there is a state of the world (ω_1, θ_1), where the seller is unable to recoup the cost of their initial investment. This is a typical hold-up problem. The contingency in which the seller fails to recoup costs may jeopardize the entire contract. Even under the ex-ante expected welfare maximizing initial contract (q_0, c_0) the buyer and seller may not agree to the contract.

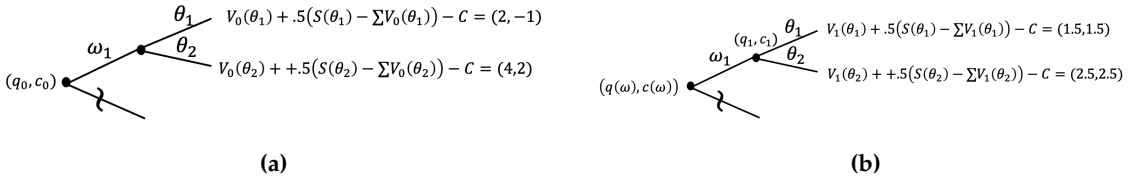


Figure C.2: *Numerical Example: (a) Explicit Prespecified Contract (q_0, c_0); (b) U.C.C.'s "Reasonable" Default $(q_d(\omega), c_d(\omega))$*

Compare the case in Figure C.2a to the case represented in Figure C.2b, in which the buyer and seller leave the delivery terms unspecified, so that the U.C.C.'s "reasonable" default will be enforced as a function of the verifiable information, $(q(\omega_1), c(\omega_1))$. When the ω-contingent default contract reigns, the reasonable default results in the buyer and seller

having valuations $V_1(\theta_1) = (1, 2)$ and $V_1(\theta_1) = (0, 1)$. When the default is renegotiated, the buyer and seller's payoffs are shown in Figure C.2b.

Note that the default adjusts when ω is realized to ensure that the division of the surplus net of investment costs is evenly distributed in both unverifiable states of the world $\{\theta_1, \theta_2\}$. This aligns with the "would have wanted" views of the role of the default in legal theory. The buyer and seller have equal bargaining weights and thus, would be expected to write contracts which evenly split the surplus. Furthermore, the "reasonable" default ensures that there are no states where the seller is unable to recoup their costs.

To summarize, in this example any explicit initial contract the parties could agree to (q_0, c_0) would be suboptimal in an ex-ante indescribable but ex-post verifiable (ω_1) state of the world. The government can play a valuable role in such cases. The government has the power to commit to enforcing a "reasonable" contract after the true state of the world is revealed, to the extent that the state is verifiable. This commitment allows the contracting parties to avoid the disaster case for the seller (ω_1, θ_1), which might otherwise jeopardize the contract.

C.2 Proofs

Proposition 1.

See Maskin and Moore (1999), Theorem 2.

Proposition 2.

We take as a starting point the reduced form game shown in Table 3.1 and replicated below:

Table C.1: *Direct mechanism for implementing (q_θ, c_θ) in state θ.*

	$\hat{\theta}_f = \theta_l$	$\hat{\theta}_f = \theta_h$
$\hat{\theta}_w = \theta_l$	(q_l, c_l)	(q_{lh}, c_{lh})
$\hat{\theta}_w = \theta_h$	(q_{hl}, c_{hl})	(q_h, c_h)

Utilities from strategy θ_h, θ_l will depend on who deviated. Specifically, if w deviates in state θ_l, their payoff will be

$$u_w(q_{hl}; \theta_l) + \delta \left(\sum_i u_i(q_l, \theta_l) - \sum_i u_i(q_{hl}, \theta_l) \right) + c_{hl}$$

If f deviated in state θ_h, their payoff will be

$$u_f(q_{hl}; \theta_h) + (1-\delta) \left(\sum_i u_i(q_h, \theta_h) - \sum_i u_i(q_{hl}, \theta_h) \right) - c_{hl}$$

In order for w and f to not deviate in these cases, the following two inequalities must hold:

$$(1-\delta)\left(u_w(q_{hl}, \theta_l) - u_w(q_l, \theta_l)\right) - \delta\left(u_f(hl, \theta_l) - u_f(q_l, \theta_l)\right) \leq c_l - c_{hl} \tag{C.1}$$

$$\delta\left(u_f(q_{hl}, \theta_h) - u_f(q_h, \theta_h)\right) - (1-\delta)\left(u_w(q_{hl}, \theta_h) - u_w(q_h, \theta_h)\right) \leq c_{hl} - c_h \tag{C.2}$$

Define $\underline{c_{hl}}$ to be the smallest value of c_{hl} such that (C.1) holds. That is,

$$\underline{c_{hl}} = c_l - (1-\delta)\left(u_w(q_{hl}, \theta_l) - u_w(q_l, \theta_l)\right) + \delta\left(u_f(q_{hl}, \theta_l) - u_f(q_l, \theta_l)\right).$$

We can then plug this expression into equation C.2 to get

$$(1-\delta)[\Delta(u_w, q_{hl}, \theta_h) - \Delta(u_w, q_{hl}, \theta_l)] - \delta[\Delta(u_f, q_{hl}, \theta_h) - \Delta(u_f, q_{hl}, \theta_l)] \geq c_l - c_h \tag{C.3}$$

where $\Delta(u, q, \theta) \equiv u(q_\theta, \theta) - u(q, \theta)$. We can do a similar calculation for the other off diagonal entry resulting in (3.5) found in the text. Next note that we can satisfy both (C.3) as well as the analogous condition for deviating in state θ_h:

$$(1-\delta)[\Delta(u_w, q_{lh}, \theta_h) - \Delta(u_w, q_{lh}, \theta_l)] - \delta[\Delta(u_f, q_{lh}, \theta_h) - \Delta(u_f, q_{lh}, \theta_l)] \leq c_l - c_h \tag{C.4}$$

by choosing $q_d = q_{hl} = q_{lh}$. This ensures that both (C.3) and (C.4) hold with equality.

Lastly, we want to show that any implementable mechanism can be implemented by default delegation. It is sufficient to show that any mechanism which implements a specific difference in transfer $c_l - c_h$, can be implemented by choosing $q_d = q_{hl} = q_{lh}$. Beginning from Proposition 1, implementability requires finding q_{lh} and q_{hl} which satisfy equations C.3 and C.4. If u_w and u_f are smooth, then it must be the case that if there exist q_{lh} and q_{hl} which satisfy equations C.3 and C.4, then there must exist $q \in [q_{lh}, q_{hl}]$, which satisfy them with equality. This is the default in the optimal default delegation mechanism.

Corollary 1.

We have shown that when u_w and u_f are continuous, if the first best is implementable, then there exists q_d such that

$$\begin{aligned}(1-\delta)[(u_w(q_h,\theta_h) - u_w(q_d,\theta_h)) - (u_w(q_l,\theta_l) - u_w(q_d,\theta_l))] \\ -\delta[(u_f(q_h,\theta_h) - u_f(q_d,\theta_h)) - (u_f(q_l,\theta_l) - u_f(q_d,\theta_l))] \\ =c_l - c_h\end{aligned} \tag{C.5}$$

where we've expanded terms in equation (3.6). We begin the proof by assuming that the condition from the corollary is met such that

$$\left|(1-\delta)\frac{\partial^2 u_w(q,\theta)}{\partial q \partial \theta} - \delta\frac{\partial^2 u_f(q,\theta)}{\partial q \partial \theta}\right| > x$$

where $x \in \mathbb{R}_+$. Furthermore, we will prove this for the specific case where

$$(1-\delta)\frac{\partial^2 u_w(q,\theta)}{\partial q \partial \theta} - \delta\frac{\partial^2 u_f(q,\theta)}{\partial q \partial \theta} > x \tag{C.6}$$

noting that the symmetric argument will hold in the opposite case. We can rewrite equation (C.5), above as

$$(1-\delta)\int_{\theta_l}^{\theta_h}\int_{q_d}^{q_h}\frac{\partial^2 u_w(q,\theta)}{\partial q\partial\theta}dqd\theta-\delta\int_{\theta_l}^{\theta_h}\int_{q_d}^{q_h}\frac{\partial^2 u_f(q,\theta)}{\partial q\partial\theta}dqd\theta$$
$$+(1-\delta)\left(u_w(q_h,\theta_l)-u_w(q_l,\theta_l)\right)-\delta\left(u_f(q_h,\theta_l)-u_f(q_l,\theta_l)\right)$$
$$=c_l-c_h$$

where we have used the fundamental theorem of calculus to substitute

$$\int_{\theta_l}^{\theta_h}\int_{q_d}^{q_h}\frac{\partial^2 u_i(q,\theta)}{\partial q\partial\theta}dqd\theta=u_i(q_h,\theta_h)-u_i(q_d,\theta_h)-\left(u_i(q_h,\theta_l)-u_i(q_d,\theta_l)\right).$$

Define $(1-\delta)\left(u_w(q_h,\theta_l)-u_w(q_l,\theta_l)\right)-\delta\left(u_f(q_h,\theta_l)-u_f(q_l,\theta_l)\right)=\kappa$, with $\kappa\in\mathbb{N}_+$. An outcome is implementable if there exists q_d which satisfies

$$(1-\delta)\int_{\theta_l}^{\theta_h}\int_{q_d}^{q_h}\frac{\partial^2 u_w(q,\theta)}{\partial q\partial\theta}dqd\theta-\delta\int_{\theta_l}^{\theta_h}\int_{q_d}^{q_h}\frac{\partial^2 u_f(q,\theta)}{\partial q\partial\theta}dqd\theta=c_l-c_h-\kappa \tag{C.7}$$

Now using our assumption (C.6) we know that

$$(1-\delta)\frac{\partial^2 u_w(q,\theta)}{\partial q\partial\theta}-\delta\frac{\partial^2 u_f(q,\theta)}{\partial q\partial\theta}dqd\theta>x$$

This means that beginning from $q_d=q_h$ and decreasing the default we can achieve any positive value of $c_l-c_h-\kappa$. By increasing the default we can achieve any negative value of $c_l-c_h-\kappa$. Thus, we have proven that when equation C.6 holds and $\gamma=0$ such that the first best aligns with the agent negotiated outcome, the first best is implementable. The same argument holds changing the inequality in equation C.6.

Corollary 2.

The result stems directly from the proof for Corollary 1. Note that as long as the condition in corollary 1 is met such that

$$\frac{\partial^2 U_w}{\partial q\partial\theta}=b>0\quad\text{and}\quad\frac{\partial^2 U_f}{\partial q\partial\theta}=-a<0$$

with $b>a$, then decreasing δ will increase

$$(1-\delta)\frac{\partial^2 u_w(q,\theta)}{\partial q \partial \theta} - \delta \frac{\partial^2 u_f(q,\theta)}{\partial q \partial \theta} dq d\theta$$

That for a given q_d, the left hand side of Equation C.7 is decreasing in δ. This proves the result as the left hand side of Equation C.7 is also strictly increasing in q_d. Therefore, a given transfer can be implemented with a lower level of q_d if δ is also lower.

Proposition 4

We continue to assume that agents Nash bargain and therefore the agents choose (q_θ, c_θ) satisfying:

$$\max_{q_\theta \in \mathcal{Q}, c_\theta \in \mathbb{R}_+} (U_w(q_\theta, c_\theta; \theta) - U_w(d;\theta))^\delta (U_f(q_\theta, c_\theta;\theta) - U_f(d;\theta))^{1-\delta} \tag{C.8}$$

Note that because the transfers are unrestricted, bargaining will be constrained efficient. The agents will choose q_θ to maximize total surplus and then choose the transfers in order to satisfy (C.8).

Another way to see this is through a contradiction. Assume that agents have chosen $q_{\theta'}$ in state θ. This implies:

$$\begin{aligned} &(U_w(q_{\theta'}, c_{\theta'};\theta) - U_w(d;\theta))^\delta (U_f(q_{\theta'}, c_{\theta'};\theta) - U_f(d;\theta))^{1-\delta} \\ &> (U_w(q_\theta, c_\theta;\theta) - U_w(d;\theta))^\delta (U_f(q_\theta, c_\theta;\theta) - U_f(d;\theta))^{1-\delta} \end{aligned} \tag{(*)}$$

Next, posit that

$$u_w(q_\theta,\theta) + u_f(q_\theta,\theta) > u_w(q_{\theta'},\theta) + u_f(q_{\theta'},\theta), \tag{C.9}$$

i.e. the agents' joint surplus from q_θ in state θ is higher than their joint surplus from $q_{\theta'}$ in state θ.

We can show that if (C.9) holds, then $q_{\theta'}$ cannot satisfy $(*)$ and therefore would not satisfy equation (C.8). To see this, note that if (C.9) holds then there must be a transfer c_θ, which satisfies

$$U_f(q_\theta, c_\theta; \theta) - U_f(d; \theta) \geq U_f(q_{\theta'}, c_{\theta'}; \theta) - U_f(d; \theta)$$

and

$$U_w(q_\theta, c_\theta; \theta) - U_w(d; \theta) \geq U_w(q_{\theta'}, c_{\theta'}; \theta) - U_w(d; \theta).$$

Furthermore, one of the two inequalities must be strict, and so we have a contradiction to $(*)$ and therefore (C.8). The agents would not choose $q_{\theta'}$ in state θ as long as (C.9) holds.

Note that this is simply an outcome of bargaining being constrained efficient. It is also useful to note that because in this case we have $\beta = 0$ the default does not play a role in achieving the first best as the regulator is unconcerned about the resulting transfers.

Corollary 3

This corollary follows directly from corollary 1. Because any set of transfers is implementable with the condition in corollary 1, then there exists a default which can replicated the desired transfers without having to restrict the set of enforceable contracts.

Proposition 5

From the proof of Proposition 5, the agents will only choose the corresponding quality if the corresponding conditions are satisfied. Note that if this is the case then we transpose into the identical situation as Proposition 2. Therefore, the same argument applies.

Proposition 6

This a mutli-state extension of Proposition 5 and the proof follows directly.

Corollary 4

This corollary is proven in the text.

Proposition 7

This proposition is derived in Appendix Section C.5 5.2.

Proposition 8

This proposition is derived in Appendix Section C.5 5.3.

C.3 Max-Min Social Welfare Functions

We show that when a regulator has a maxmin objective function, the results characterizing first-best implementability with default delegation apply when there are more than two states.

A max-min objective function takes the form

$$\max_{\mathcal{Q},d} \min_{\theta} \mathbb{E}[\text{SWF}(q^*, c^*; \theta)] \tag{C.10}$$

subject to incentive constraints. We say a SWF is *max-min implementable* if the first-best outcome corresponding to the worst-case state can be implemented in Nash equilibrium and any refinement.

Suppose θ_k is the state that delivers the worst welfare for the principal, i.e.

$$\theta_k \in \arg\min_{\theta} \text{SWF}(q, c; \theta).$$

Then, the principal can partition the state space Θ into cells $\{\theta_1, \theta_2, \dots\}, \{\theta_k\}$. The states that are not θ_k can be treated as one, and the principal can offer a delegation set $\{(q_{-k}, c_{-k}), (q_k, c_k)\}$ and a use a default that satisfies (3.6). Although social welfare will not be "first-best" in all states that are not k, it can be first best in state k.

Consider the following example:

Example 5 *Consider the following specification of preferences:*

- $u_w(q, \theta) = -(q - \theta)^2$ *(worker's bliss point $q = \theta$),*

- $u_f(q) = -q^2$ *(firm's bliss point $q = 0$), and*
- $\Theta = \{\theta_l, \theta_m, \theta_h\} = \{0, \frac{1}{2}, 1\}$.

With these preferences, the regulator's welfare is minimized in the high state θ_h. So, a max-min regulator cares only about maximizing welfare in state θ_h. In such a case, the regulator can design the direct mechanism in Table C.2. The first-best outcome in the high state is $(q_h, c_h) = (\frac{\theta}{2}, \frac{1}{4})$.

Table C.2: *Default delegation with max-min regulator* ($|\Theta| = 3$).

	$\hat{\theta}_f \in \{\theta_l, \theta_m\}$	$\hat{\theta}_f = \theta_h$
$\hat{\theta}_w \in \{\theta_l, \theta_m\}$	(q_{-h}, c_{-h})	(q_d, c_d)
$\hat{\theta}_w = \theta_h$	(q_d, c_d)	(q_h, c_h)

The conditions for implementation of (q_{-h}, c_{-h}) in states θ_l and θ_m and (q_h, c_h) in state θ_h are

$$(1-\delta)[\Delta(u_w, q_h, q_d, \theta_h) - \Delta(u_w, q_{-h}, q_d, \theta_{-h})] - \delta[\Delta(u_f, q_h, q_d, \theta_h) - \Delta(u_f, q_{-h}, q_d, \theta_{-h})] \geq c_{-h} - c_h \tag{C.11}$$

and

$$(1-\delta)[\Delta(u_w, q_h, q_d, \theta_h) - \Delta(u_w, q_{-h}, q_d, \theta_{-h})] - \delta[\Delta(u_f, q_h, q_d, \theta_h) - \Delta(u_f, q_{-h}, q_d, \theta_{-h})] \leq c_{-h} - c_h \tag{C.12}$$

for $\theta_{-h} \in \{\theta_l, \theta_m\}$. The regulator here has many degrees of freedom to choose (q_{-h}, c_{-h}) and (q_d, c_d) so that the constraints above are both satisfied with equality for $\theta_{-h} \in \{\theta_l, \theta_m\}$, i.e.

$$(1-\delta)[\Delta(u_w, q_h, q_d, \theta_h) - \Delta(u_w, q_{-h}, q_d, \theta_{-h})] - \delta[\Delta(u_f, q_h, q_d, \theta_h) - \Delta(u_f, q_{-h}, q_d, \theta_{-h})] = c_{-h} - c_h. \tag{C.13}$$

This example shows that even with $|\Theta| > 2$, the two state case offers valuable insight. For instance, when a regulator has max-min preferences, the results from the two state case (under some further conditions) show us that the regulator can implement the first best in the worst-case state.

More generally, we can learn from the two state cases in order to understand the relative importance of particular states. Note that in the case where the regulator only has concerns about efficiency and equity, the optimal transfers lead to equal surplus in a particular state. Thus, the deviation from optimal can be measured by the difference between the optimal transfer and the transfer induced by a particular default. Thus, we can imagine solving for q_d, c_d by using Equation 3.12 and then plugging in that value into equation Equation 3.13 in order to solve for the transfer outcome $\tilde{c}_2$ where the tilde denotes that it is not necessarily optimal. You could similarly perform this exercise to extract $\tilde{c}_1$. The losses from optimizing across states 1 and 3 would be smaller than the losses from optimizing across 2 and 3 if $|\tilde{c}_2 - c_2| < |\tilde{c}_1 - c_1|$.

C.4 Extensions

C.4.1 Inequity penalty in SWF.

Our model assumes that the social welfare function the regulator wants to maximize takes on a particular functional form. The efficiency and externality terms in the social welfare function are standard—efficiency is total surplus between the worker and the firm ($U_w + U_f$) and the externality is captured by γU_r, where U_r is some well-behaved function. The equity term, on the other hand, is less standard. We model the regulator's preference for equality as a quadratic penalty $-\beta(U_w - U_f)^2$, which may appear extreme in its implications. In particular, it suggests that for any positive β, the first-best contract results in a perfectly equal distribution of the surplus in all states of the world.

Although this assumption may not always align with how regulators and lawmakers think about equity in practice, there are two key reasons why it is a useful assumption for understanding the theoretical limits faced by the regulator.

First, the equality quadratic penalty term is really a stand-in for a broader class of weighted quadratic penalty terms: it captures the idea that in the first-best contract, the regulator wants *a particular* distribution of the surplus. The fact that the regulator wants

an equal distribution of the surplus is besides the point—the analysis would be largely unchanged if the SWF featured a weighted penalty term with

$$-\beta(\alpha U_w - (1-\alpha)U_f)^2$$

for $\alpha \in (0,1)$. For example, with $\alpha = .75$, the first-best would always feature a 75-25 split of the surplus.

Second, our focus on a particular distribution of surplus is the most restrictive assumption we could make about the regulator's equality preferences. In practice, lawmakers tolerate inequality up to point. So the first-best distributions of surplus often are not a single division, but instead feature a range of possible ways of dividing the surplus. For example, a lawmaker may only think inequality is objectionable if the worker's utility is less than (greater than) some fraction $\bar{k}$ ($\underline{k}$) of the firm's utility. That is, the regulator's inequity penalty may be

$$\beta(\mathbf{1}[\underline{k} < U_w/U_f < \bar{k}])$$

which implies that the first-best is a set of contracts, rather than a singleton. Since we are interested here in understanding the *limits* on implementation, we focus on singleton case which will be the most limited case. If the first-best featured a range of possible distribution splits, first-best would be "easier" to implement. In this sense, our assumption about the quadratic penalty term is a limiting case that offers bounds on implementability of more realistic equity preferences.

C.4.2 Exogenous income

The firm and worker may enter the contracting environment with exogenous income. Exogenous income has consequences for the regulator's distributional preferences, and also may affect limited liability constraints, which we have not discussed. With exogenous income, the worker and firm preferences are given by

$$\textbf{Firm: } U_f(q,c;\theta) = y_f + u_f(q;\theta) - c \quad \textbf{Worker: } U_w(q,c;\theta) = y_w + u_w(q;\theta) + c,$$

where y_i represents an exogenous and separable component of preferences for agent $i \in \{w, f\}$. In practice, y_w may represent workers' wages pinned down by an outside option whereas y_f is the firm's total revenue net of these wages.

When there is a regulator with preferences over efficiency and equality, these exogenous parameters will simply effect the water-level of the transfers. An increase (decrease) in y_f or a decrease (increase) in y_w will necessitate a higher (lower) transfer in all states. This does not effect the regulator's ability to implement the first-best however, as the first best only depends on the state-dependence components of outcomes. Thus, we have the following proposition.

Proposition 9 *Assume a regulator is maximizing social welfare which has efficiency, equity and externality components. Any exogenous and separable components of utility which are state-independent and do not depend on quality, only affect the optimal mechanism by adjusting the default transfer. This does not impact implementability.*

C.5 Application Derivations and Discussion

Assume that the firm can provide additional wage in order to compensate for the risk at the job by paying workers c.

$$U_w = WS(c,q;\theta) = w - (q-\theta)^2 + c$$
$$U_f = \Pi(c,q) = R - q^2 - c$$

The regulator wants to maximize

$$U_p = WS + \Pi - \beta(WS - \Pi)^2 + \gamma q$$

so that the regulator cares about the firm and worker surplus, the distribution and a term which captures an externality or social value associated with quality.

Summarizing:

- w: baseline pay independent of occupational risk
- θ: captures the optimal level of safety for the consumer and affects the tradeoff between additional wages versus higher safety
- R: revenue of the firm net of baseline pay to worker
- c: compensation tied to safety concerns
- q: level of safety measures
- $G(\theta)$: Prior over the state. Throughout we will assume $\theta \sim U(0,1)$

C.5.1 Regulator Problem

The regulator can choose a default (q_d, c_d) and cannot prevent renegotiation. However, they can impose a minimum, $\underline{q}$ and maximum quality level, $\overline{q}$. Without any constraints on the quality level we know that the firm and the consumer will negotiate until $q_\theta = \frac{\theta}{2}$. This allows us to divvy up the state space into three intervals. We also know that the transfer will be the outcome of the renegotiation from the default and will be given by

$$c_\theta = c(q_\theta, c_d, q_d; \theta) = c_d + (1-\delta)(-(q_d - \theta)^2 + (q_\theta - \theta)^2)) - \delta(-q_d^2 + q_\theta^2))$$

$$c_\theta = c(q_\theta, c_d, q_d; \theta) = c_d + (1-\delta)(q_\theta^2 - q_d^2 + 2\theta(q_d - q_\theta)) - \delta(-q_d^2 + q_\theta^2))$$

$$c_\theta = c(q_\theta, c_d, q_d; \theta) = c_d + (1-2\delta)q_\theta^2 - (1-2\delta)q_d^2 + 2(1-\delta)\theta(q_d - q_\theta)$$

Binding minimum: $\theta < 2\underline{q} \rightarrow q_\theta = \underline{q}$

Unconstrained Interval: $2\underline{q} < \theta < 2\overline{q} \rightarrow q_\theta = \frac{\theta}{2}$

Binding Maximum: $\theta > 2\overline{q} \rightarrow q_\theta = \overline{q}$

Why Interval Delegation?

Alonso and Matouschek (2008) characterizes conditions under which interval delegation is optimal. Their conditions apply in the case of delegation to a single agent, and have to do with the difference between the agent and the principal's optimal decisions.

A key object in their analysis is the *backward bias*, defined as

$$T(\theta) = G(\theta)(q_\theta - E[q^*(z)|z \leq \theta])$$

where $q^*(z)$ is the regulator's optimal quality choice conditional on bargaining. They show that if $T''(\theta) > 0$ in the relevant range of θ, then the regulator will choose interval delegation.

Proposition 10 (Alonso and Matouschek (2008),Proposition 2 (i)) *Let $\mathcal{Q}^*$ be an optimal delegation set. Then if $T(\theta)$ is strictly convex then Q^* contains either no decision, one decision, or an interval of decisions.*

We can apply this result in our example to confirm that the regulator would choose interval delegation. In our example, the backward bias is

$$T(\theta) = \theta\left(\frac{\theta}{2} - \frac{1}{\theta}\int_0^\theta (q^*(z))dz\right) = \frac{\theta^2}{2} - \int_0^\theta q^*(z)dz$$

and its second derivative is

$$T''(\theta) = 1 - \frac{\partial q^*(\theta)}{\partial \theta}. \tag{C.14}$$

Next we need to solve for $q^*(\theta)$, the regulator's optimal quality outcome conditional on the state. Taking the default as given, $q^*(\theta)$ is defined by the problem

$$q^*(\theta) = \arg\max_q \left[w + R - (q-\theta)^2 - q^2 - \beta(-(q-\theta)^2 + q^2 + 2c)^2 + \gamma q\right]. \tag{C.15}$$

The term c in the maximand above is in fact a function of q, δ, and q_d. In this set up, we are treating δ and q_d as exogenous, and can rewrite c explicitly,

$$c = c_d + (1-2\delta)q^2 - (1-2\delta)q_d^2 + 2(1-\delta)\theta(q_d - q).$$

Now plugging this expression for c into (C.15) and taking the first order condition, we get

$$-4q^*(\theta)+2\theta+\gamma-2\beta\left(w-R+2c_\theta-(q^*(\theta)-\theta)^2+q^*(\theta)^2\right)\left(2\theta+4\left[(1-2\delta)q^*(\theta)-\theta(1-\delta)\right]\right)=0. \tag{C.16}$$

To make this expression concrete, consider the simplest case when $\delta = .5$. In this case the first order condition (C.16) simplifies to

$$q^*(\theta)=\frac{2\theta+\gamma}{4}.$$

That is, the regulator's optimal q is the agent-optimal q shifted by $\frac{\gamma}{4}$. Substituting $q^*(\theta)$ into the equation for the second derivative of the backward bias (C.14) yields

$$T''(\theta)=\frac{1}{2}.$$

By proposition 2 in Alonso and Matouschek (2008), the fact that $T''(\theta) > 0$ implies that the regulator's optimal delegation set is either no decision, one decision or an interval.

Next consider the case where $\delta = 0$. In this case, the regulator's first-order condition in (C.16) simplifies to

$$-4q^*(\theta)+2\theta+\gamma-2\beta\left(w-R+2c_\theta+2q^*(\theta)\theta-\theta^2\right)\left(4q^*(\theta)-2\theta\right)=0.$$

Furthermore, note that the default transfer will always cancel the initial inequality such that we can consider $w = R = 0$ without loss of generality. Substituting in the transfer we get

$$-4q^*(\theta)+2\theta+\gamma-2\beta\left(2\left(c_d+q^*(\theta)^2-q_d^2+2\theta(q_d-q^*(\theta))\right)+2q^*(\theta)\theta-\theta^2\right)\left(4q^*(\theta)-2\theta\right)=0$$

Implicitly differentiating yields:

$$-4\frac{\partial q^*(\theta)}{\partial\theta}+2-2\beta\left(4\frac{\partial q^*(\theta)}{\partial\theta}-2\right)\left(2c_\theta+2q^*(\theta)\theta-\theta^2\right)$$
$$-2\beta(4q^*(\theta)-2\theta)\left(4q^*(\theta)\frac{\partial q^*(\theta)}{\partial\theta}+2(q_d-q^*(\theta))-2\theta\frac{\partial q^*(\theta)}{\partial\theta}\right)$$

Simplifying, yields

$$\frac{\partial q^*(\theta)}{\partial\theta}=\frac{1+2\beta(2c_\theta-\theta^2+2\theta q_d-4q^*(\theta)(q_d-q^*(\theta)))}{2+4\beta(2c_\theta+2q^*(\theta)\theta-\theta^2+(2q^*(\theta)-\theta)^2)}$$

As long as this expression is less than 1, the second derivative of the backward bias in (C.14) will be positive. We can note a few circumstances under which this expression is less than one. First, when the equity parameter β is small, this expression is less than 1. Note that this also will be the case when the changes in equity are relatively small from figure 3.3 that the impact on inequity is lower in the middle than at the extremes which will make this condition more likely to hold.

A similar analysis can be undertaken for the case of $\delta = 1$.

Solving the Regulator's Problem

The regulator problem is then given by

$$\max_{\{(c_d,q_d,\underline{q},\overline{q})} \int_{\underline{\theta}}^{2\underline{q}} \left(w + R - (\underline{q} - \theta)^2 - \underline{q}^2 - \beta \left(w - R + 2c(\underline{q}, c_d, q_d; \theta) - (\underline{q} - \theta)^2 + \underline{q}^2 \right)^2 + \gamma \underline{q} \right) dG(\theta) +$$
$$\int_{2\underline{q}}^{2\overline{q}} \left(w + R - \frac{\theta^2}{2} - \beta \left(w - R + 2c(\theta/2, c_d, q_d; \theta) \right)^2 + \gamma \frac{\theta}{2} \right) dG(\theta) +$$
$$\int_{2\overline{q}}^{\overline{\theta}} \left(w + R - (\overline{q} - \theta)^2 - \overline{q}^2 - \beta \left(w - R + 2c(\overline{q}, c_d, q_d; \theta) - (\overline{q} - \theta)^2 + \overline{q}^2 \right)^2 + \gamma \overline{q} \right) dG(\theta)$$

Subject to the constraints:

$$2\underline{q} \geq \underline{\theta}$$
$$2\overline{q} \leq \overline{\theta}$$
$$2\overline{q} \geq \underline{\theta}$$
$$\underline{q} \leq q_d \leq \overline{q}$$

It is useful here to totally differentiate $c(q_\theta, c_d, q_d; \theta)$

$$dc_\theta = dc_d + 2\left[(2\delta - 1)q_d + \theta(1 - \delta)\right] dq_d + 2\left[(1 - 2\delta)q_\theta - \theta(1 - \delta)\right] dq_\theta$$

Now we can take first order conditions.

$$\frac{\partial V}{\partial \underline{q}} = \int_{\underline{\theta}}^{2\underline{q}} \left(-4\underline{q} + 2\theta + \gamma - 2\beta\left(w - R + 2c_\theta - (\underline{q} - \theta)^2 + \underline{q}^2\right)\left(2\theta + 4\left[(1-2\delta)\underline{q} - \theta(1-\delta)\right]\right)\right) dG(\theta)$$

$$\frac{\partial V}{\partial \overline{q}} = \int_{2\overline{q}}^{\overline{\theta}} \left(-4\overline{q} + 2\theta + \gamma - 2\beta\left(w - R + 2c_\theta - (\overline{q} - \theta)^2 + \overline{q}^2\right)\left(2\theta + 4\left[(1-2\delta)\overline{q} - \theta(1-\delta)\right]\right)\right) dG(\theta)$$

$$\begin{aligned}\frac{\partial V}{\partial c_d} = 0 = \frac{w-R}{2} + \\ &\int_{\underline{\theta}}^{2\underline{q}} \left(c_d + (1-2\delta)\underline{q}^2 - (1-2\delta)q_d^2 + 2(1-\delta)\theta(q_d - \underline{q}) - \frac{\theta^2}{2} + \underline{q}\theta\right) dG(\theta) + \\ &\int_{2\underline{q}}^{2\overline{q}} \left(c_d + (1-2\delta)\frac{\theta^2}{4} - (1-2\delta)q_d^2 + 2(1-\delta)\theta\left(q_d - \frac{\theta}{2}\right)\right) dG(\theta) + \\ &\int_{2\overline{q}}^{\overline{\theta}} \left(c_d + (1-2\delta)\overline{q}^2 - (1-2\delta)q_d^2 + 2(1-\delta)\theta(q_d - \overline{q})) - \frac{\theta^2}{2} + \overline{q}\theta\right) dG(\theta)\end{aligned}$$

$$\begin{aligned}\frac{\partial V}{\partial q_d} = 0 = &\int_{\underline{\theta}}^{2\underline{q}} \left(w - R + 2c_\theta - (\underline{q} - \theta)^2 + \underline{q}^2\right)\left[(2\delta - 1)q_d + \theta(1-\delta)\right] dG(\theta) + \\ &\int_{2\underline{q}}^{2\overline{q}} \left(w - R + 2c_\theta\right)\left[(2\delta - 1)q_d + \theta(1-\delta)\right] dG(\theta) + \\ &\int_{2\overline{q}}^{\overline{\theta}} \left(w - R + 2c_\theta - (\overline{q} - \theta)^2 + \overline{q}^2\right)\left[(2\delta - 1)q_d + \theta(1-\delta)\right] dG(\theta)\end{aligned}$$

C.5.2 Equal Bargaining $\delta = .5$

It is useful to start by considering the case where the firm and worker have equal bargaining positions. In this case they will split any surplus generated from renegotiation equally. Importantly, the regulator's equality term aims to equate worker surplus and firm profits. Thus, the bargaining conditions are aligned with the regulator's incentives as far as equity is concerned.

We also know that the regulator's externality term favors high quality. Combined this means that the regulator would never want to implement a maximum quality level.

However, it would implement a minimum quality level. When $\delta = .5$ the FOC's simplify to

$$\frac{\partial V}{\partial \underline{q}} = \int_{\underline{\theta}}^{2\underline{q}} \left(-4\underline{q} + 2\theta + \gamma\right) dG(\theta)$$

$$\frac{\partial V}{\partial \overline{q}} = \int_{2\overline{q}}^{\overline{\theta}} \left(-4\overline{q} + 2\theta + \gamma\right) dG(\theta)$$

$$\frac{\partial V}{\partial c_d} = 0 = \frac{w - R}{2} + \int_{\underline{\theta}}^{\overline{\theta}} \left(c_d + \theta q_d - \frac{\theta^2}{2}\right) dG(\theta)$$

$$\frac{\partial V}{\partial q_d} = 0 = \int_{\underline{\theta}}^{\overline{\theta}} \theta \left(w - R + 2\left(c_d + \theta\left(q_d - \frac{\theta}{2}\right)\right)\right) dG(\theta)$$

Note that as expected there is no dependence on β. We can also solve explicitly the regulator's problem:

$$\underline{q} = \frac{\gamma}{2}$$

$$\overline{q} = \frac{\overline{\theta}}{2} = \frac{1}{2}$$
$$c_d = \frac{R - w}{2} - \mu_\theta q_d + \frac{\mu_{\theta^2}}{2}$$

$$q_d = \frac{1}{2}\frac{\mu_{\theta^3} - \mu_{\theta^2}\mu_\theta}{\mu_{\theta^2} - \mu_\theta \mu_\theta}$$

Note that the term inside the integral for $\frac{\partial V}{\partial \overline{q}}$ is always positive and thus, the maximum is at a corner solution and thus, does not bind. We also see that $\underline{q}$ is increasing in γ as that is the only way to ensure a higher quality level when there is renegotiation. Lastly we have that the default transfer with our assumed distribution is $c_d = \frac{R-w}{2} - \frac{1}{12}$ and $q_d = \frac{1}{2}$. To interpret this we note that the default transfer's first job is to equate the initial surplus $R + w$ between the two parties. Then we see that the quality default is the expected optimal for the worker. Thus, the regulator is paying the worker who has state dependent preferences in

terms of the quality whereas the firm is being paid in terms of a negative default transfer beyond the initial redistribution. Interestingly the presence of the minimum does not affect the default quality and transfer. This is because the only role of the defaults are to ensure equality. Since the two parties always equally divide the surplus, the realized outcomes do not affect equality, only the relationship between the default and the state.

It is interesting to consider an alternative distribution. Below are the results when $g(\theta) = 1 + 2(\theta - .5) \quad \text{for} \quad \theta \in (0,1)$. With this distribution we will now get a higher minimum quality level and a different default quality and transfer. Specifically,

$$\underline{q} = \frac{3\gamma}{4}$$
$$\overline{q} = \frac{1}{2}$$
$$c_d = \frac{R-w}{2} - \mu_\theta q_d + \frac{\mu_{\theta^2}}{2} = \frac{R-w}{2} - \frac{3}{20}$$
$$q_d = \frac{1}{2}\frac{\mu_{\theta^3} - \mu_{\theta^2}\mu_\theta}{\mu_{\theta^2} - \mu_\theta\mu_\theta} = \frac{3}{5}$$

Note that the default quality has gone up, the default transfer has gone down and the responsiveness to externalities has gone up. All of these make sense given that the benefit of a high default is higher now that the typical state is higher. Furthermore, we can compare the default quality to the expected worker optimal as before. The expected value of the state is now $2/3$, which means that although the mean value increasing has pushed up the default quality, the skewed distribution has made it so that the default quality is no longer as high as the workers expected optimal.

C.5.3 Firm Power $\delta = 0$

The most interesting case and the case that gives us intuition for the realm where we may expect the regulator to be operating is when the firm has all of the bargaining power. From our discussion in implementation theory, we know that this is the state where the regulator is in the best position to implement something approximating first best because the worker is more sensitive to the true state.

Substituting $\delta = 0$ into the FOCs yields

$$\frac{\partial V}{\partial \underline{q}} = \int_{\underline{\theta}}^{2\underline{q}} \left(-4\underline{q} + 2\theta + \gamma - 2\beta\left(w - R + 2c_\theta - (\underline{q} - \theta)^2 + \underline{q}^2\right)\left(4\underline{q} - 2\theta\right)\right) dG(\theta)$$

$$\frac{\partial V}{\partial \overline{q}} = \int_{2\overline{q}}^{\overline{\theta}} \left(-4\overline{q} + 2\theta + \gamma - 2\beta\left(w - R + 2c_\theta - (\overline{q} - \theta)^2 + \overline{q}^2\right)\left(4\overline{q} - 2\theta\right)\right) dG(\theta)$$

For the solver it is useful to rewrite these equations:

$$\frac{\partial V}{\partial \underline{q}} = (2\underline{q})\left(\gamma - 2\underline{q} - 4\beta\underline{q}\left(2c_d + \frac{8}{3}q_d\underline{q} - 2q_d^2 + w - R\right)\right) = 0$$

First thing to note is that for weakly positive $\underline{q}$ and $\gamma = 0$, this equation holds when $\beta = 2$. You can also rewrite it defining $\underline{d} = 2\underline{q}$.

$$\frac{\partial V}{\partial \underline{q}} = \underline{d}\left(\gamma - \underline{d} - 2\beta\underline{d}\left(2c_d + \frac{4}{3}q_d\underline{d} - 2q_d^2 + w - R\right)\right) = 0$$

$$\frac{\partial V}{\partial \overline{q}} = 0 = (1 - 2\overline{q})(1 + \gamma - 2\overline{q} + \frac{\beta}{3}(1 - 2\overline{q})(-3 + 12c_d + 4(4 - 3q_d)q_d + 4\overline{q}(4q_d - 3) + 6w - 6R))$$

Similarly, defining $\overline{d} = 1 - 2\overline{q}$

$$\frac{\partial V}{\partial \overline{q}} = 0 = \overline{d}\left(\gamma + \overline{d} + 2\beta\overline{d}\left(2c_d - \frac{4}{3}q_d\overline{d} - 2q_d^2 + 4q_d - \frac{3}{2} + w - R\right)\right)$$

We can solve these two explicitly in the case wehere $\gamma = 0$ and $\overline{d}$ and $\underline{d}$ are weakly positive.

$$\underline{d} = \frac{3}{4q_d}\left(\frac{-1}{2\beta} - 2c_d + 2q_d^2 + w - R\right)$$

$$\overline{d} = \frac{\left(\frac{-1}{2\beta} - 2c_d + 2q_d^2 + w - R + 4q_d - \frac{3}{2}\right)}{\frac{4}{3}q_d - 1}$$

We can also substitute in for $c_d = \frac{R-w}{2} + \frac{3\mu_{\theta 2}}{4} + q_d^2 - 2q_d\mu_\theta$.

$$\underline{d} = \frac{\frac{-1}{2\beta} - \frac{3\mu_{\theta^2}}{2} + 4q_d\mu_\theta}{\frac{4}{3}q_d}$$

$$\overline{d} = \frac{\left(\frac{-1}{2\beta} - \frac{3\mu_{\theta^2}}{2} + 4q_d(1+\mu_\theta) - \frac{3}{2}\right)}{\frac{4}{3}q_d - 1}$$

We can then differentiate these with respect to q_d

$$d\underline{d} = \frac{\frac{16}{3}q_d\mu_\theta - \frac{4}{3}\left(\frac{-1}{2\beta} - \frac{3\mu_{\theta^2}}{2} + 4q_d\mu_\theta\right)}{(4q_d/3)^2}dq_d$$

$$d\overline{d} = \frac{4(q_d+\mu_\theta)(\frac{4}{3}q_d - 1) - \frac{4}{3}\left(\frac{-1}{2\beta} - \frac{3\mu_{\theta^2}}{2} + 4q_d(1+\mu_\theta) - \frac{3}{2}\right)}{(\frac{4}{3}q_d - 1)^2}dq_d$$

At $\beta = 2$, $q_d = 3/8$. These two equations simplify to

$$d\underline{d} = 4dq_d$$

$$d\overline{d} = -7dq_d$$

$$\begin{aligned}\frac{\partial V}{\partial c_d} = 0 = \frac{w-R}{2} + \\ \int_{\underline{\theta}}^{2\underline{q}} \left(c_d + \underline{q}^2 - q_d^2 + 2\theta(q_d - \underline{q}) - \frac{\theta^2}{2} + \underline{q}\theta\right) dG(\theta) + \\ \int_{2\underline{q}}^{2\overline{q}} \left(c_d + \frac{\theta^2}{4} - q_d^2 + 2\theta\left(q_d - \frac{\theta}{2}\right)\right) dG(\theta) + \\ \int_{2\overline{q}}^{\overline{\theta}} \left(c_d + \overline{q}^2 - q_d^2 + 2\theta(q_d - \overline{q}) - \frac{\theta^2}{2} + \overline{q}\theta\right) dG(\theta)\end{aligned}$$

Or

$$\frac{\partial V}{\partial c_d} = 0 = \frac{w-R}{2} + c_d - \frac{1}{6} + \frac{1}{6}\left(6(1-q_d)q_d + \overline{q}(-3 + 6\overline{q} - 4\overline{q}^2) + 4q_l^3\right)$$

$$\frac{\partial V}{\partial q_d} = 0 = \int_{\underline{\theta}}^{2\underline{q}} \left(w - R + 2c_\theta - (\underline{q} - \theta)^2 + \underline{q}^2\right) \left[-q_d + \theta\right] dG(\theta) + \int_{2\underline{q}}^{2\overline{q}} \left(w - R + 2c_\theta\right) \left[-q_d + \theta\right] dG(\theta) + \int_{2\overline{q}}^{\overline{\theta}} \left(w - R + 2c_\theta - (\overline{q} - \theta)^2 + \overline{q}^2\right) \left[-q_d + \theta\right] dG(\theta)$$

or

$$\frac{\partial V}{\partial q_d} = 0 = -3 + \left(c_d + \frac{w - R}{2}\right)(12 - 24q_d) - 36q_d^2 + 24q_d^3 - 4\overline{q}(2 - 3\overline{q} + 2\overline{q}^3) + 8\underline{q}^4 + 4q_d(5 + \overline{q}(3 - 6\overline{q} + 4\overline{q}^2) - 4$$

Furthermore, we can substitute in the value for c_d to get:

$$q_d = \frac{1 + 2\overline{q} - 8(1 - \overline{q})\overline{q}^3 + 8(1 - \underline{q})\underline{q}^3}{4}$$

$$4dq_d = \left(2 - 8(3 - 4\overline{q})\overline{q}^2\right) d\overline{q} + 8\left(3 - 4\underline{q}\right) \underline{q}^2 d\underline{q}$$

As before it is useful to solve these FOCs for when there is neither a minimum nor maximum. For instance, in the limit as $\gamma \to 0$ and $\beta < 2$. In this case,

$$c_d = \frac{R - w}{2} + \frac{3\mu_{\theta^2}}{4} + q_d^2 - 2q_d\mu_\theta = \frac{R - w}{2} + \frac{1}{64}$$

$$q_d = \frac{3}{8}\frac{\mu_{\theta^3} - \mu_{\theta^2}\mu_\theta}{\mu_{\theta^2} - \mu_\theta\mu_\theta} = \frac{3}{8}$$

Note that the default quality is lower when the workers have a worse bargaining position. This because any renegotiation away from the default will result in higher surplus for the firm relative to the worker. Thus, in order to reduce inequality across states it is useful to reduce the extent of renegotiation and adjust the transfer.

Similarly, we can solve for $\underline{q}$ when $\beta \to 0$ resulting in $\underline{q} = \frac{\gamma}{2}$, which unsurprisingly is the same as in the equal bargaining case. As in either case the minimum is defined by efficiency and externality concerns and are independent of equality.

It is also useful to consider how the minimum adjusts with inequality. In order to gain some intuition consider we can evaluate $\frac{\partial^2 V}{\partial \underline{q} \partial \beta}$ at $\beta = 0$

$$\frac{\partial^2 V}{\partial \underline{q} \partial \beta}|_{\beta=0} = -4\underline{q}\left(2c_d + \frac{8}{3}q_d\underline{q} - 2q_d^2 + w - R\right)$$

We can substitute in the values for c_d and q_d to get

$$\frac{\partial^2 V}{\partial \underline{q} \partial \beta}|_{\beta=0} = -4\underline{q}\left(\frac{3}{2}\mu_{\theta^2} - 4q_d\mu_\theta + \frac{8}{3}q_d\underline{q}\right) = -4\underline{q}\left(\frac{1}{2} - \frac{3}{4} + \underline{q}\right)$$

This means that starting from $\underline{q} = 0$, the minimum is increasing in β. In general we expect that the interval over which the the firm and worker are able to choose quality is decreasing in concerns over equality. Note that with the firm making take it or leave it offers to the worker,

C.5.4 Worker Power $\delta = 1$

Another special case is when the workers have all of the bargaining power. The difficulty here from the standpoint of the regulator is that the firm does not have state dependent preferences. This highlights the importance of state dependence as highlighted in the implementation discussion above. Thus, the default which is used to smooth the necessary transfer across states can no longer perform this function. We can see this directly in the first order conditions by plugging in $\delta = 1$.

$$c_\theta = c(q_\theta, c_d, q_d; \theta) = c_d - q_\theta^2 + q_d^2$$

Note that the transfer does not depend on the state except through the level of quality.

$$\frac{\partial V}{\partial \underline{q}} = \int_{\underline{\theta}}^{2\underline{q}} \left(-4\underline{q} + 2\theta + \gamma - 2\beta\left(w - R + 2c_\theta - (\underline{q} - \theta)^2 + \underline{q}^2\right)\left(-2\underline{q}\right)\right) dG(\theta)$$

$$\frac{\partial V}{\partial \overline{q}} = \int_{2\overline{q}}^{\overline{\theta}} \left(-4\overline{q} + 2\theta + \gamma - 2\beta\left(w - R + 2c_\theta - (\overline{q} - \theta)^2 + \overline{q}^2\right)\left(-2\overline{q}\right)\right) dG(\theta)$$

$$\frac{\partial V}{\partial c_d} = 0 = \frac{w-R}{2} + \int_{\underline{\theta}}^{2\underline{q}} \left(c_d - \underline{q}^2 + q_d^2 - \frac{\theta^2}{2} + \underline{q}\theta \right) dG(\theta) + \int_{2\underline{q}}^{2\overline{q}} \left(c_d - \frac{\theta^2}{4} + q_d^2 \right) dG(\theta) + \int_{2\overline{q}}^{\overline{\theta}} \left(c_d - \overline{q}^2 + q_d^2 - \frac{\theta^2}{2} + \overline{q}\theta \right) dG(\theta)$$

$$\frac{\partial V}{\partial q_d} = 0 = \int_{\underline{\theta}}^{2\underline{q}} \left(w - R + 2\left(c_d - \underline{q}^2 + q_d^2 \right) - \theta^2 + 2\underline{q}\theta \right) [-q_d]\, dG(\theta) + \int_{2\underline{q}}^{2\overline{q}} \left(w - R + 2\left(c_d - \frac{\theta^2}{4} + q_d^2 \right) \right) [-q_d]\, dG(\theta) + \int_{2\overline{q}}^{\overline{\theta}} \left(w - R + 2\left(c_d - \overline{q}^2 + q_d^2 \right) - \theta^2 + 2\overline{q}\theta \right) [-q_d]\, dG(\theta)$$

Note that because q_d is just a constant we can eliminate it from the last expression which shows that when the workers have all of the bargaining power, the transfer and quality level are only jointly determined. We can show this by rewriting the last equation

$$\frac{\partial V}{\partial q_d} = 0 = \frac{w-R}{2} + \int_{\underline{\theta}}^{2\underline{q}} \left(c_d - \underline{q}^2 + q_d^2 - \frac{\theta^2}{2} + \underline{q}\theta \right) dG(\theta) + \int_{2\underline{q}}^{2\overline{q}} \left(c_d - \frac{\theta^2}{4} + q_d^2 \right) dG(\theta) + \int_{2\overline{q}}^{\overline{\theta}} \left(c_d - \overline{q}^2 + q_d^2 - \frac{\theta^2}{2} + \overline{q}\theta \right) dG(\theta)$$

which is the same as the first order condition for c_d. If the maximum and minimum are not binding this simply implies that $\frac{R-w}{2} - c_d - q_d^2 = \frac{\mu_{\theta^2}}{4} = -\frac{1}{12}$, which is the expected losses that the workers will experience from the optimal level of quality. Thus, the firm will always receive $\frac{R+w}{2} - \frac{1}{12}$ as will the workers in expectation. If the minimum is binding, the workers will be made better off relative to the firm in low states of the world and thus, the default will become more favorable to the firm. Similarly, if there is a maximum, then the maximum

will make the firm relatively better off compared to the workers and thus, the default will be less favorable to the firm.

Furthermore, we can define the default value of the firm as $F_d = -c_d - q_d^2$ and reconsider the regulator problem as choosing $\underline{q}, \overline{q}, F_d$. Rewriting and simplifying the FOCs:

$$\frac{\partial V}{\partial \underline{q}} = \int_{\underline{\theta}}^{2\underline{q}} \left(-4\underline{q} + 2\theta + \gamma + 4\underline{q}\beta \left(w - R + 2\left(-F_d - \underline{q}^2\right) + 2\underline{q}\theta - \theta^2\right)\right) dG(\theta)$$

$$\frac{\partial V}{\partial \overline{q}} = \int_{2\overline{q}}^{\overline{\theta}} \left(-4\overline{q} + 2\theta + \gamma + 4\overline{q}\beta \left(w - R + 2\left(-F_d - \overline{q}^2\right) - (\overline{q} - \theta)^2 + \overline{q}^2\right)\right) dG(\theta)$$

$$\begin{aligned}\frac{\partial V}{\partial F_d} = 0 = \frac{w - R}{2} + \\ \int_{\underline{\theta}}^{2\underline{q}} \left(-F_d - \underline{q}^2 - \frac{\theta^2}{2} + \underline{q}\theta\right) dG(\theta) + \\ \int_{2\underline{q}}^{2\overline{q}} \left(-F_d - \frac{\theta^2}{4}\right) dG(\theta) + \\ \int_{2\overline{q}}^{\overline{\theta}} \left(-F_d - \overline{q}^2 - \frac{\theta^2}{2} + \overline{q}\theta\right) dG(\theta)\end{aligned}$$

Solving for F_d yields:

$$F_d = \frac{w - R}{2} - \frac{4}{3}\underline{q}^3 - \frac{2}{3}(\overline{q}^3 - \underline{q}^3) + \frac{4}{3}\overline{q}^3 - \overline{q}^2 - \frac{1}{6} + \overline{q}/2$$

$$F_d = \frac{w - R}{2} - \frac{2}{3}\underline{q}^3 + \frac{2}{3}\overline{q}^3 - \overline{q}^2 - \frac{1}{6} + \overline{q}/2$$

It is important here to note that $\frac{\partial F_d}{\partial \underline{q}} < 0$ and $\frac{\partial F_d}{\partial \overline{q}} > 0$. These signs are intuitive when we consider how the worker's utility changes in the range where either the minimum or maximum is binding. In either case, the worker's utility is decreasing the tighter is the constraint because they get all of the surplus from renegotiation. Thus, when the minimum $\underline{q}$ raises the worker's position deteriorates which induces an optimal reduction in the firm

surplus. Similarly when $\overline{q}$ decreases, and thus, the constraint binds more tightly, the firm value F_d must decrease.

However, this immediately suggests that the maximum must never bind because we know that the worker's utility is always below the firm's in the high state. In the low state $\theta = 0$, the worker receives its optimal quality as well as receiving the the maximum transfer from the firm. In the highest state, the firm is made the worst off from renegotiation which minimizes the value the workers receive from renegotiation. In this case the maximum can only increase inequality.

In order to see this we can look directly at the inequality term. We know that the worker's utility is given by $\frac{w+R}{2} - (q_\theta - \theta)^2 - F_d - q_\theta^2$ Meanwhile the firm receives $\frac{w+R}{2} + F_d$. The difference then is given by

$$Eq = W - F = -(q_\theta - \theta)^2 - 2F_d - q_\theta^2$$

With renegotiation, we can plug in $\theta/2$ and get

$$Eq = -\frac{\theta^2}{2} - 2F_d$$

This is a decreasing function of the state in the relevant range $\theta > 0$. Thus, in order to minimize the deviation it must be that $F_d > -\frac{\overline{\theta}^2}{4}$, so that $Eq < 0$ when $\theta = \overline{\theta}$. Now consider decreasing q_θ where $\theta = \overline{\theta}$. In this case,

$$\frac{\partial Eq}{\partial q_\theta} = -2(q_\theta - \theta) - 2q_\theta$$

Note that this is positive when $q_\theta < \frac{\theta}{2}$. Thus, any decrease in the maximum will lead to a decrease in Eq which given that Eq is negative implies greater inequality. Thus, we know that $\overline{q} = .5$ in this setting.

Now we can solve for the minimum by substituting in $F_d = \frac{w-R}{2} - \frac{2}{3}\underline{q}^3 - \frac{1}{12}$.

$$\frac{\partial V}{\partial \underline{q}} = \int_{\underline{\theta}}^{2\underline{q}} \left(\underbrace{-4\underline{q} + 2\theta}_{\textbf{efficiency}} + \underbrace{\gamma}_{\textbf{externality}} + \underbrace{4\underline{q}\beta\left(\frac{4}{3}\underline{q}^3 + \frac{1}{6} - 2\underline{q}^2 + 2\underline{q}\theta - \theta^2\right)}_{\textbf{equality}} \right) dG(\theta)$$

Note here that it is clear that losses due to efficiency will reduce the value of the minimum, the externality will push the minimum up and the equality term will also lead to greater minimums.