Essays on Credit Frictions, Market Expansion, and Strategic Team Production

Benjamin Tengelsen
Dec 19, 2018

Submitted to the Tepper School of Business
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
at
Carnegie Mellon University

Doctoral Committee:

Sevin Yeltekin (Chair)
Laurence Ales
Christopher Telmer
Ariel Zetlin-Jones

For Ray Meyers - Bozeman's best calculus teacher

ACKNOWLEDGEMENTS

I have benefited from the help and assistance of many people while working on this dissertation. I'm fortunate to have had my wife Laura as a support and confidant throughout my time as a student. Our children Nash, Samson, and Elaine have also given me inspiration, perspective, and an unmistakable urgency to finish. I'm also grateful for the support I've received from my parents and siblings.

I received many hours of world-class coaching from an excellent dissertation committee, comprised of Sevin Yeltekin, Laurence Ales, Chris Telmer, and Ariel Zetlin-Jones. Their advice and perspective were central to the development of these ideas. I am also thankful for the informal mentorship of current and former CMU faculty: Kate Anderson, Brian Routledge, Chris Sleet, and Fallaw Sowell. I'm similarly grateful to my former BYU mentors Rick Evans and Kerk Phillips for helping me succeed even after I graduated from BYU. Finally, I thank Lawrence Rapp and Laila Lee for their wonderful administrative help.

The first and last chapters in this dissertation represent joint efforts with other economists. I thank my coauthors Emilio Bisetti, Nicolas Petrosky-Nadeau, Etienne Wasmer, and Ariel Zetlin-Jones for their extended collaboration, mentorship, and friendship. I'm especially grateful for Nicolas Petrosky-Nadeau for allowing me to work at the Federal Reserve Bank of San Fransisco for extended periods of time.

For many helpful conversations and for making the student years enjoyable, I thank all of my fellow Tepper students and especially my office neighbors Emilio Bisetti, Leah Clark, Hakk Özdenören, Eungsik Kim, Maxime Roy, and Alex Schiller. I also thank my friends outside of Tepper - Nate Bringhurst, Hayden Cardiff, Chase Coleman, Bill Morales, and

Ryan Morrison for befriending my family during our time in Pittsburgh. Finally, I thank my manager at Wayfair, Zhenyu Lai, for granting me the flexibility to finish this dissertation while working full-time this past year. ’

ABSTRACT

Essays on Credit Frictions, Market Expansion, and Strategic Team Production

by

Benjamin Tengelsen

Chair: Sevin Yeltekin

The first chapter, jointly authored with Nicolas Petrosky-Nadeau and Etienne Wasmer, studies the relationship between credit markets and labor markets over the business cycle. We explicitly categorize US quarters between 1953 and 2017 as being "recession", "normal", or "expansion" based on the deviation of unemployment from its long-run trends. We then examine how various credit-market measures correlate with unemployment in the following quarters. We find changes in the credit market have correlations with future unemployment that vary dramatically with the initial state of the economy. We then show that the same patterns of state-dependency exist in a model with search-frictional credit and labor markets. After calibrating the model to match key labor and credit-market moments, we estimate impulse response functions and find the impact of any adverse shock on unemployment to be meaningful only under certain initial conditions. We also find that while unemployment is about 1.6 times more responsive to productivity shocks than credit-market shocks, the response of the credit spread is about even between productivity and credit-market shocks.

In the second chapter, I examine several instances where the removal of geographic barriers caused increased competition between formerly isolated firms, resulting in fewer firms and a more concentrated market. Notable instances of this pattern include the US commercial

banking industry, the US retail industry in response to the advent of e-commerce, exporting firms following the removal of international trade barriers, and the US brewing industry following the adoption of national television and mass advertising. I propose a theoretical model that explicitly accounts for geographic distance and the power it grants firms to act monopolistically within their local markets. As these geographic barriers are removed over time, either gradually or suddenly, prices experience downward pressure from increased competition and upward pressure as firms exit and surviving firms inherit larger market shares. I also explore a range of parameter values that demonstrate nonlinear relationships between market size and market concentration. While market concentration is generally increasing in these settings, increased market expansion can also reduce firm output such that large firms acquire less market share in the long-run even though the number of active firms has decreased.

The final chapter, jointly authored with Emilio Bisetti and Ariel Zetlin-Jones, re-examines the importance of separation between ownership and labor in team production models that feature free riding. In such models, conventional wisdom suggests an outsider is needed to administer incentive schemes that do not balance the budget. We analyze the ability of insiders to administer such incentive schemes in a repeated team production model with free riding when they lack commitment. Specifically, we augment a standard, repeated team production model by endowing insiders with the ability to impose group punishments which occur after team outcomes are observed but before the subsequent round of production. We extend techniques from *Abreu* (1986) to characterize the entire set of perfect-public equilibrium payoffs and find that insiders are capable of enforcing welfare enhancing group punishments when they are sufficiently patient.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

CHAPTER I

Credit and Labor-Market Frictions over the Business Cycle

1.1 Introduction

What is the relationship between credit and labor markets and how does it vary over the business cycle? In this paper we document empirical evidence suggesting there is a strong degree of state dependence in the relationship between fluctuations in credit-market spreads and unemployment in the following quarters. During recessions, the estimated response of unemployment to an increase in credit spreads is many times larger than during normal times. This would suggest that credit-market shocks normally play a modest role in business cycles, but are capable of playing a substantial role if the economy has already begun to slow down.

This business cycle asymmetry arises naturally in a model with search-frictional labor and financial markets, and we use such a model to provide an interpretable lens on our empirical findings. In a typical search model, the probability of finding a match varies with the relative size of the unmatched parties (e.g. unemployed individuals, firms without creditors), which can vary significantly over the business cycle. Consequently, the impact of an adverse shock will depend on the aggregate state of the economy and its corresponding matching probabilities. In the case of the labor market, the magnitude of the response to

an adverse shock is positively correlated with the unemployment rate. Similarly, the credit market is increasingly sensitive to additional shocks when there are many unmatched firms seeking creditors. In both cases, tighter matching markets increase the elasticity of job creation to shocks. The asymmetric effect of shocks in this economy can thus originate in either the labor or credit market, as well as both simultaneously.

The model builds on *Wasmer and Weil* (2004) and *Petrosky-Nadeau and Wasmer* (2012). Firms form matches with financial institutions in a search-frictional credit market in order to expand productive capacity, which for simplicity we refer to as a job. Financial institutions provide funds to a firm when the job is open and searching for a worker in the search frictional labor market, and receive a share of the profit flow generated when the job is filled. Prices in the credit and labor market are determined by Nash bargaining. The model nests the canonical Diamond-Mortensen-Pissarides (DMP) as a special case when the credit market is removed.

We consider two sources of business cycle fluctuations, shocks to labor productivity and shocks to the cost of credit-market search for financial institutions. Productivity shocks affect firms directly as part of the production function. Credit-market shocks affect the effort financial institutions put into searching in the financial market. An adverse shock reduces a financial institution's search effort, making it harder for the firm to increase its production capacity. Moreover, a negative shock increases the value of the financial institution's outside option in bargaining with the firm over the repayment. This further squeezes profits away from the firm, depressing job creation and making the economy more vulnerable to additional shocks.

The dynamic properties of the model are first illustrated by its ability to match a collection of state-dependent moments. We attempt a novel calibration strategy in which moments are computed for recessions and normal periods, and the model parameters are tuned until the model simulations match volatility moments in both states. The model's state-dependent properties are also evident through its theoretical impulse responses to productivity and

credit-market shocks at different initial conditions. The response of both unemployment and the credit-spread is negligible when the adverse shocks arrive during expansionary or even normal periods. Only when unemployment is already high and the economy is beginning from a relatively poor position does an adverse shock cause large movements in our variables of interest. We also find that during a period of high unemployment, unemployment is about 1.6 times more responsive to productivity shocks than credit-market shocks. The response of the credit spread, however, is about even between productivity and credit-market shocks.

This paper follows a long line of research into the macroeconomic consequences of financial frictions on the business cycle. Early work modeled either agency costs or problems of limited commitment in financial markets (*Bernanke et al.*, 1996, *Kiyotaki and Moore*, 1997). In *Bernanke et al.*, 1996 agency costs in lending relationships introduce a financial accelerator that amplifies business cycles. *House* (2006) shows that, in general, models of adverse selection in financial markets will either amplify or mitigate business cycles, depending on whether the friction leads to insufficient or excessive investment. Collateral constraints have been shown in some contexts to provide a powerful amplification mechanism (*Cordoba and Ripoll*, 2004), especially when a fixed resource such as land serves as collateral (*Liu et al.*, 2013).

A more recent literature has approached modeling financial markets as search markets (*Wasmer and Weil*, 2004; *Lagos and Rocheteau*, 2009; *Petrosky-Nadeau*, 2013). The model of Section 1.3 builds on the work of *Wasmer and Weil* (2004) and *Petrosky-Nadeau and Wasmer* (2012), which studies the business cycle dynamics of the labor market in the presence of search-frictional labor and financial markets. Their work establishes the efficiency properties of the model, and in particular the existence of a Hosios-type condition in the financial market which minimizes the amplifying factor of the financial friction. This work further develops the notion of financial institutions and casts the theory in a representative agent environment. In addition, it introduces shocks to financial markets very much in the spirit of *Jermann and Quadrini* (2012). *Bai* (2016) uses a Diamond-Mortensen-Pissarides with

defaultable debt to examine credit-spreads and finds that the model does well at matching key properties of the credit market.

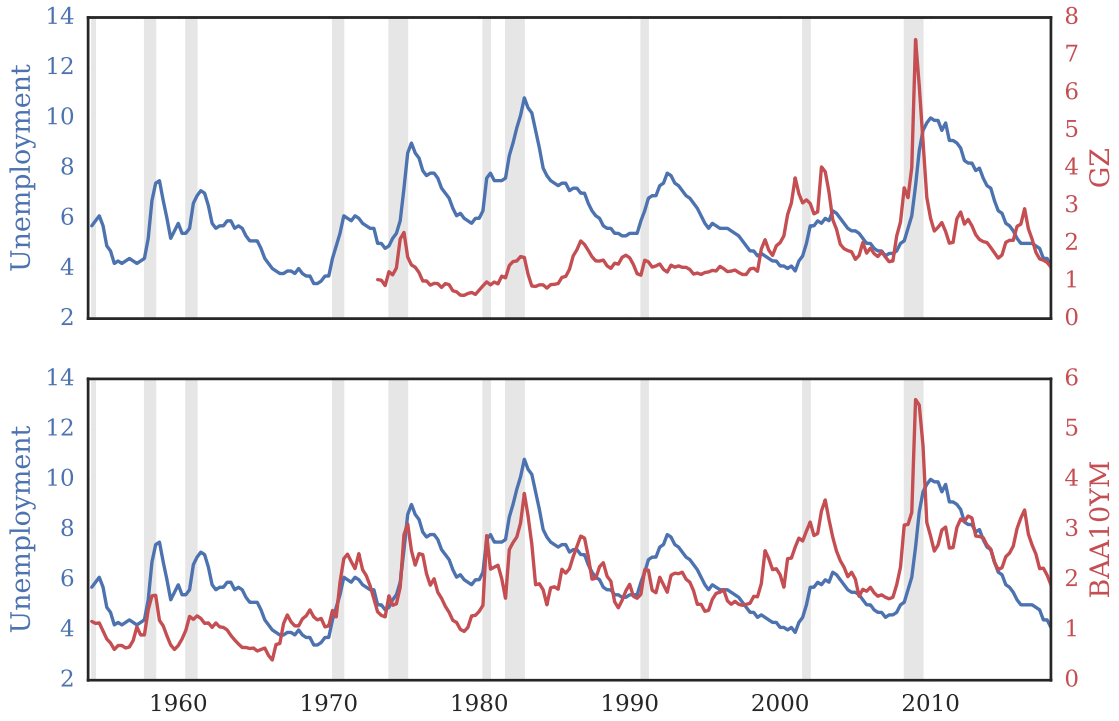
The business cycle literature notes time series asymmetries in the unemployment rate for the U.S. in work such as *Neftci* (1984). *Petrosky-Nadeau and Zhang* (2013a) show that the congestion externality of the matching function leads the search and matching model of equilibrium unemployment of DMP to generates unemployment times series with deep troughs during recessions and a degree of skewness in line with U.S. data (see also *Hairault et al.*, 2010). Our empirical evidence adopts the flexible framework developed by *Jordà* (2005). Other research focused on the asymmetric impact of fiscal shocks, implements smooth transition VARs (*Auerbach and Gorodnichenko*, 2012, 2014, *Caggiano et al.*, 2014). However, that approach estimates the impact of a shock under the assumption that the current state of the economy will endure indefinitely. In the approach we follow, the future impact of a shock accounts for the most likely state of the economy following its initial regime.

Section 1.2 presents the empirical evidence on the effects of credit-market shocks over the business cycle. Section 1.3 develops the model of the macroeconomy with search-frictional labor and financial markets, while the quantitative results are in Section 1.4. Section 1.5 concludes.

1.2 Credit-market shocks over the business cycle

We first show the asymmetry of the relationship between credit and labor markets over the business cycle within a simple regression framework. Specifically, we use credit market data, unemployment, and indicator variables for the aggregate state of the economy to estimate the response of unemployment to a change in credit market conditions. As shown in *Jordà* (2005), the coefficients of this regression trace out an empirical impulse response function as the forecast horizon is extended. Moreover, with this framework we can compute the impulse responses of unemployment conditional on the initial state of the economy.

Figure 1.1: Time Series of Unemployment and Credit Spreads*



*Grey areas indicate NBER recessions. GZ spread is obtained from *Gilchrist and Zakrajšek* (2012) and is a composite of a broad range of outstanding senior unsecured bonds. BAA is an investment bond that acts as an index for all bonds given a BAA rating by Moody's Investor Service.

1.2.1 Data and Econometric Framework

Our economic outcome of interest is the unemployment rate $\mathcal{U}$. All time series are at a quarterly frequency. We use two data sources for measuring credit-market conditions, the spread between BAA corporate bonds and 10 year treasury notes and the *Gilchrist and Zakrajšek* (2012) "GZ credit spread."[1] The BAA-10 Year spread is especially appropriate to consider in relation to unemployment, as it compensates for default risk in addition to non-default factors such as liquidity risk which correlate less with unemployment (*Bai* (2016)). Both credit-market series are plotted along side the unemployment rate over the period

[1]The GZ spread is constructed using micro-data on a broad range of corporate bonds, and is a better predictor of changes in the real economy than other popular corporate bond spreads. *Gilchrist and Zakrajšek* (2012) show the GZ spread to be useful in forecasting unemployment over short horizons. Our forecasts differ from theirs in that we include variables that allow for business cycle asymmetries.

1953:II to 2017:I in Figure 1.1.[2] Data for the GZ spread are only available from 1973:II to 2017:IV. Grey, shaded, areas indicate NBER recessions dates.

The BAA spread and unemployment track each other closely with a correlation coefficient of 0.57. Generally speaking, both series demonstrate sharp increases during economic downturns and demonstrate markedly less volatility during normal times. The GZ spread also spikes during recessions along with the unemployment rate, but does not measure a strong statistical correlation with unemployment. The contemporaneous correlation coefficient between the two series is .01.

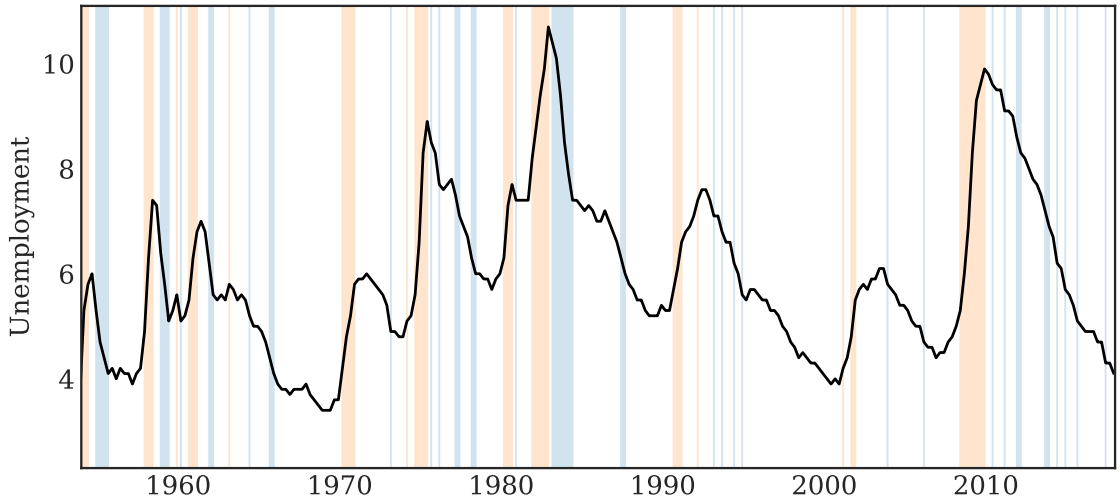
In order to measure the correlation between changes in the credit market to subsequent changes in the unemployment rate, we estimate the following regression:

$$\mathcal{U}_{t+h} \quad = \quad \beta_0 + \beta_R(L)R_t + \beta_D(L)D_t + \beta_{DR}(L)DR_t + \beta_X(L)X_t + \varepsilon_{t+h}. \tag{1.1}$$

The dependent variable, $\mathcal{U}_{t+h}$, is the h-step-ahead forecast of the U.S. unemployment rate with $h > 0$. On the right hand side, the R_t is a measure of credit-market activity, and D_t is a matrix of dummy variables indicating whether the economy is in a period of high, normal, or low unemployment. We refer to these periods, respectively, as "recession", "normal", and "expansion" states of the economy. Specifically, D_t includes two dummy variables D_{EXP} and D_{REC} to indicate whether the economy is in an expansion or a recession. The normal state occurs when D_{EXP} and D_{REC} are both zero. For ease of interpretation, we prefer to interact our credit-market variable with a small number of discrete states rather than a continuous variable, but similar results are obtained. These interaction terms are contained in the matrix DR_t, and are key regressors in our analysis. The coefficients on these interaction terms indicate whether or not credit markets move symmetrically with unemployment over the business cycle. The matrix X_t contains additional control variables and summarizes all additional information available at time t. The lag operator denotes how many historical

[2]Unemployment rate for the civilian population over the age of 16, published by the Bureau of Labor Statistics (BLS), based on the Current Population Survey (CPS).

Figure 1.2: Discrete Economic States as Determined by First-differenced U_t



Red shaded regions denote recessions while blue shaded regions denote expansions. Recessions are defined to be periods where first-differenced unemployment is above its 80th percentile. Expansions are defined to be periods where first-differenced unemployment is below its 20th percentile. The data range from 1953:Q3 to 2017:Q4.

values of each variable are included as additional controls. Finally, ε_{t+h} is our h-step-ahead forecast error.

The state of the economy is determined by the a first-difference time series of unemployment which we write as $\tilde{\mathcal{U}}$. The economy is said to be in a recession when $\tilde{\mathcal{U}}$ exceeds a threshold amount $\overline{\mathcal{U}}$. In the context of equation (1.1), $D_{REC,t} = 1$ when $\tilde{\mathcal{U}}_t > \overline{\mathcal{U}}$. The threshold $\overline{\mathcal{U}}$ is chosen so that the economy is in a recession 20 percent of our sample, which is only slightly more frequent than recessions as dated by the NBER and is consistent with the definition of a recession as a period of rapidly increasing unemployment. Similarly, we define a lower threshold $\underline{\mathcal{U}}$ for expansion states, setting $D_{EXP,t} = 1$ when $\tilde{\mathcal{U}}_t < \underline{\mathcal{U}}$. This threshold is selected such that an expansion also occurs in 20% of periods. Figure A.1 shows the time series for U_t with shaded regions indicating the state of the economy. Our rule of thumb for characterizing the state of the economy captures the well-known recessions in the post-war era. Changing $\overline{\mathcal{U}}$ and $\underline{\mathcal{U}}$ by small amounts (plus or minus three percentiles of $\tilde{\mathcal{U}}$) has no substantial impact on our results.

Matrix X_t contains three control variables, including the period t unemployment rate, the vacancy to unemployment ratio θ_t, and labor productivity x_t. The time series for vacancies is taken from *Petrosky-Nadeau and Zhang* (2013b). Labor productivity is measured as real output per person for all non-farm business sectors, and is measured as a percent deviation from a long-run trend which we identify via an HP-filter with $\lambda = 1600$. We choose the optimal lags in each regression via the AIC and BIC selection criteria, using the smaller of the two.

Our choice of methodology merits some comment as we depart from a frequently used approach in measuring business cycle asymmetries. Several studies use smoothly varying weights to indicate the state of the economy between regimes (see *Auerbach and Gorodnichenko*, 2012, 2014, *Caggiano et al.*, 2014). However, smooth transition VARs assume the state of the economy to be permanent, and ignore the probability that the economy moves to another regime in the future. This assumption will obviously bias the resulting impulse response functions. Under our approach, the regime is allowed to vary according to the average path of the economy, moving away from an initial regime to another. This approach is more realistic over longer horizons. Our approach, as described in *Jordà*, 2005, is also more robust to erroneous specifications and handles nonlinearities with greater accuracy. [3]

1.2.2 Empirical results

Equation (1.1) is estimated by ordinary least squares. The coefficients of interest, those on R_t and its interaction terms at different forecast horizons are reported in Table 1.1. Panel A reports the results for the BAA-10 year spread, and panel B reports the results using the GZ spread. The coefficients can be interpreted as the level response of unemployment to a 1 point increase in the credit-market spread. The first row indicates the response when the economy is in a "normal" state. The second and third rows report the additional impact on

[3] *Auerbach and Gorodnichenko* (2014) combine these methods by augmenting local projection regressions with smooth transition weights. However, we prefer dummy variables for their transparency and ease of interpretation.

Table 1.1: The Relationship between Credit Markets and Future Unemployment at Different Forecast Horizons and in Different Economic States

Panel A: Credit market series $R = BAA - 10\ year$ spread
Time sample: 1953:II-2017:IV

Forecast horizon:	h=1	h=3	h=6	h=9
R	0.325 **	0.416*	0.203*	0.156
	(0.128)	(0.251)	(0.390)	(0.258)
R_REC	0.190	0.440***	0.551**	0.529
	(0.148)	(0.249)	(0.292)	(0.251)
R_EXP	-0.097*	-0.261	-0.228	-0.039
	(0.086)	(0.157)	(0.222)	(0.263)
Constant	0.309*	0.519	0.641	0.671
	(0.475)	(0.995)	(1.427)	(1.842)
Observations	253	251	248	245

Panel B: Credit market series $R = GZ$ spread
Time sample: 1973:I-2017:IV

Forecast horizon:	h=1	h=3	h=6	h=9
R	0.282***	0.381	0.175	0.165
	(0.099)	(0.244)	(0.349)	(0.369)
R_REC	0.153	0.295	0.349	0.261
	(0.120)	(0.285)	(0.323)	(0.349)
R_EXP	0.096	0.042	-0.176	-0.524
	(0.115)	(0.207)	(0.276)	(0.336)
Constant	0.276	0.155	0.653	-1.516
	(0.497)	(.976)	(1.463)	(1.655)
Observations	173	171	168	165

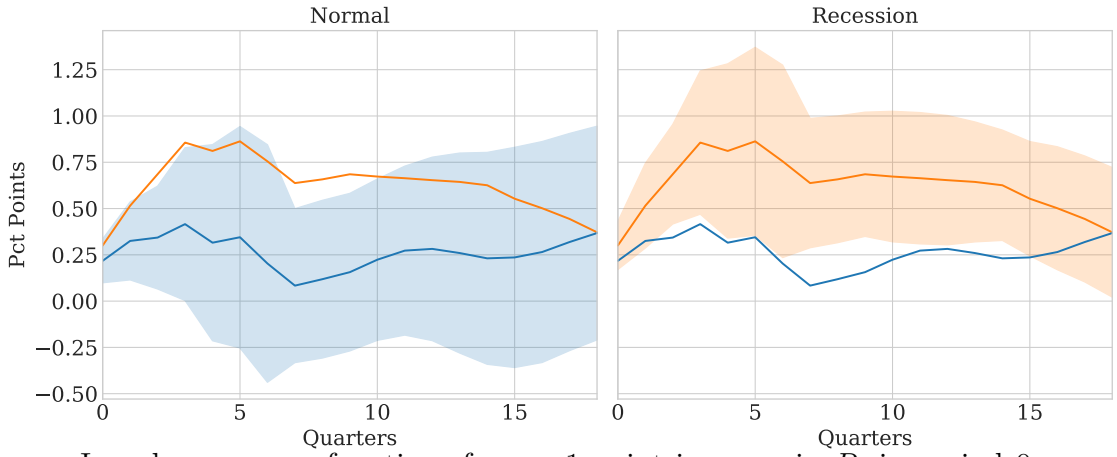
Standard errors in parentheses. ***: p<0.01, **: p<0.05,: * p<0.1

unemployment when the economy is in a recession or expansion, respectively. The interaction terms $R \times D_{rec}$ and $R \times D_{exp}$ allow the relationship between unemployment and the various credit spreads to vary depending on the state of the economy.

In normal times there is a small, positive correlation between innovations to the spread and the unemployment rate. A unit increase in the BAA-10 year spread is associated with a 0.33 percentage point increase in the unemployment rate the following quarter ($h = 1$). The coefficients are positive for all values of the forecast horizon h, and peak in the sixth quarter with a 0.55 percentage point increase in unemployment. The coefficients for $R \times D_{rec}$, reported in the second row, are positive for all the forecast horizons considered, and significant for forecast horizons of 0 through 1 quarters. These coefficients suggests that credit-market shocks matter more, or are associated with more pronounced increases in unemployment, when they occur during recessions as opposed to normal times. The peak in the additional reaction in unemployment to credit market shock during a recession occurs after 6 quarters, adding 0.55 more percentage points to the unemployment rate relative to response in normal times. This additional increase by itself is more than twice the response of unemployment in normal times. The coefficient on the interaction term for expansions, $R \times D_{exp}$, is significant only for $h = 1$ and is negative, suggesting a smaller response in unemployment during expansionary periods.

The results using the GZ spread as a measure of credit-market conditions similarly demonstrate asymmetric responses in unemployment over the business cycle. The smaller response from the GZ spread is expected to a degree as the GZ spread remains fairly flat up until the mid 1990's. Another reason to expect a lesser response from the GZ spread is that the spread is an unweighted average of coorporate bond spreads including relatively riskless bonds which track closely with treasuries. The $BAA10YM$ spread, on the other hand, is focused only on relatively risky busineses which are more likely to fail during an economic downturn. The response from a GZ shock occurring in normal times peaks at 3 quarters but loses statistical significance after 2 quarters. A shock during a recession, however, has

Figure 1.3: Estimated Response of U_t to a Unit Increase in $BAA10YM$ Spread

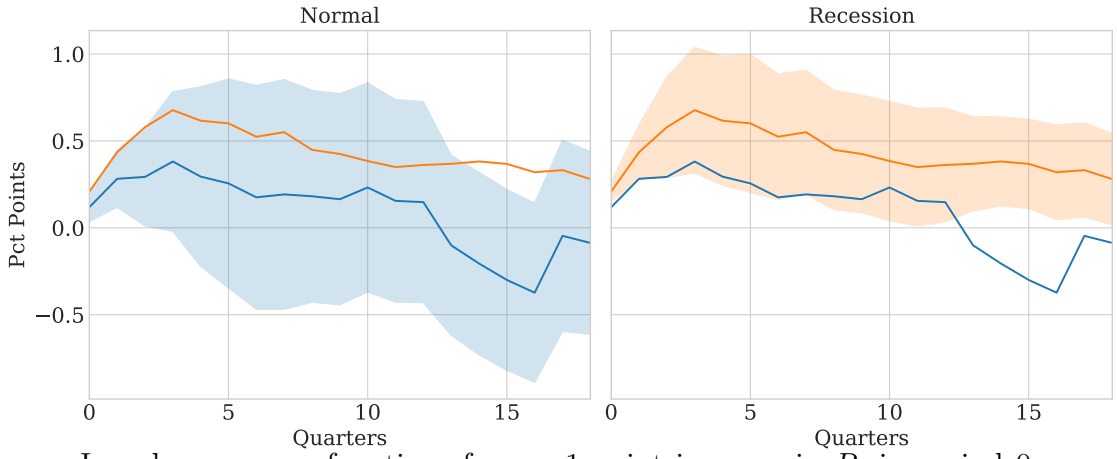


Impulse response functions from a 1 point increase in R_t in period 0. The left panel includes 90% confidence interval for the 'normal' phase of the business cycle. The right panel features confidence intervals for 'recession' phases.

a response that is statistically greater than zero for over 10 quarters. At h=3, the total response is over 75% larger in a recession than in normal times. At this forecast horizon, a unit increase in the GZ spread leads to a .67 percentage point increase in unemployment.

The coefficients estimated from (1.1) allow us to trace out the impulse response of unemployment to an innovation in the credit-market spread (see *Jordà*, 2005). The effects of a unit increase in the BAA-10 year spread on the unemployment rate are plotted in Figure 1.3 under two different scenarios. In the first case, the blue line, the economy is in a normal phase of the business cycle when the innovation occurs. In the second, the effects of the innovation to the spread when the economy is already in a recession are plotted in Figure 1.3 as the red line. When considering the BAA-10 year spread, the response of the unemployment rate in the period of the innovation in normal times is consistently less than when the economy is in a recession until $h = 12$. The impulse responses using the GZ spread are plotted in Figure 1.4. The pattern is similar to the previous case, if not more pronounced. The peak response occurs later, and the difference relative to normal times is larger. In both cases, the standard errors are large, even when individual coefficients are statistically significant, as they combine the standard errors from both coefficients (β_R and β_{DR}) as well

Figure 1.4: Estimated Response of U_t to a Unit Increase in GZ Spread



Impulse response functions from a 1 point increase in R_t in period 0. The left panel includes 90% confidence interval for the 'normal' phase of the business cycle. The right panel features confidence intervals for 'recession' phases.

as their covariance.

1.3 Model

We model an economy with search frictions in labor and credit markets, building on the work of *Wasmer and Weil* (2004) and *Petrosky-Nadeau and Wasmer* (2012). A representative household provides labor to produce output and makes current risk free bond and consumption choices. Firms produce with labor and finance their expansion efforts through a frictional financial market in which they are paired with a creditor. A creditor is an institution maximizing profits for its shareholders (the representative household) by managing a large number of credit relationships creating new credit matches.

1.3.1 Matching in financial and labor markets

In order for firms to create an additional job, they must first establish a partnership with a creditor to finance the upfront costs associated with recruiting a worker. At any point in time there are $\mathcal{N}_{ct}$ such projects searching for a creditor. On the other side of the financial

market, financial intermediaries place $\mathcal{B}_{ct}$ units of effort to seek new projects with which to be matched. Meetings in the financial market are governed by the constant returns to scale matching function $M_c(\mathcal{B}_{ct}, \mathcal{N}_{ct})$, which is increasing and concave in both arguments. We use ϕ_t the denote the ratio $\mathcal{N}_{ct}/\mathcal{B}_{ct}$, which reflects credit market tightness from the point of view of new projects. The contact rates for each side of the credit market are:

$$\begin{aligned} p_t &= \frac{M_c(\mathcal{B}_{ct}, \mathcal{N}_{ct})}{\mathcal{N}_{ct}} = p(\phi_t) \text{ with } p'(\phi_t) < 0, \\ \bar{p}_t &= \frac{M_c(\mathcal{B}_{ct}, \mathcal{N}_{ct})}{\mathcal{B}_{ct}} = \bar{p}(\phi_t) \text{ with } \bar{p}'_t(\phi_t) > 0. \end{aligned}$$

The first equation states the probability p_t of a project matching with a creditor in a unit of time is a decreasing function of credit market tightness. The second equation states the rate $\bar{p}_t$ at which a creditor matches with a project is an increasing function of the relative abundance of investment projects. The assumption of constant returns in matching implies $\bar{p}_t = \phi_t p_t$.

New positions are added to the pool of vacant jobs $\mathcal{V}_t$ in the labor market. These job vacancies are sought after by the unemployed $\mathcal{U}_t$. We normalize the labor force to 1, and consequently $\mathcal{U}_t$ also denotes the current unemployment rate. Matching in the labor market is governed by the function $M_l(\mathcal{V}_t, \mathcal{U}_t)$, which demonstrates constant returns and is increasing in all arguments. We define the ratio $\mathcal{V}_t/\mathcal{U}_t = \theta_t$ as the tightness of the labor market from the perspective of the firm. The meeting rates for each side of the labor market are:

$$\begin{aligned} q_t &= \frac{M_l(\mathcal{V}_t, \mathcal{U}_t)}{\mathcal{V}_t} = q(\theta_t) \text{ with } q'(\theta_t) < 0, \\ f_t &= \frac{M_l(\mathcal{V}_t, \mathcal{U}_t)}{\mathcal{U}_t} = f(\theta_t) \text{ with } f'(\theta_t) > 0. \end{aligned}$$

The first line states a vacancy is filled in period t with probability $q(\theta_t)$, which is decreasing in labor-market tightness. The second line states that a worker finds employment in a unit of time with probability $f(\theta_t)$. This job -finding probability is increasing in labor-market

tightness. The assumption of constant returns in matching implies $f_t = \theta_t q_t$.

1.3.2 Firms

A firm produces with linear production technology $Y_t = X_t \mathcal{N}_t$. Here, with slight abuse of notation from Section 1.2, X_t denotes productivity and is both exogenous and stochastic. The variable $\mathcal{N}_t$ is the share of workers currently engaged in production. In order to hire a worker and generate output, a firm must first create additional productive capacity which will be either vacant or filled with a worker. This requires searching for a new credit relationship in the financial market. Searching in the credit market is costly, incurring a flow cost κ_I per project. Once a firm establishes a match in the financial market, the creditor finances the costs of searching in the labor market. The creditor pays recruiting cost γ when a position is vacant, and is paid an amount Ψ_t when the position is filled and generating revenue. Each period the firm pay workers a wage W_t. Prior to the start of the next period, a deterministic share of matches in both the labor and credit markets are severed at random. The labor-market separation rate is given by $s^L \in (0,1)$. Separations in the labor market become open vacancies, but the firm-creditor match remains intact. Credit relationships separate at rate $s^C \in (0,1)$, in which case the entire position is destroyed.

Given this environment, the firm's objective is to maximize the value of its equity by choosing the amount of projects to place on financial markets, $\mathcal{N}_{ct}$:

$$S_t = \max_{\mathcal{N}_{ct}} \left[X_t \mathcal{N}_t - W_t \mathcal{N}_t - \Psi_t \mathcal{N}_t - \kappa_I \mathcal{N}_{ct} \right] + \mathbb{E}_t M_{t+1} \left[S_{t+1} \right] \tag{1.2}$$

$$\text{subject to} \qquad \mathcal{V}_t = \left(1 - s^C\right) \left[\left(1 - q(\theta_{t-1})\right) \mathcal{V}_{t-1} + s^L \mathcal{N}_{t-1} \right] + p(\phi_t) \mathcal{N}_{ct} \tag{1.3}$$

$$\mathcal{N}_{t+1} = \left(1 - s^C\right) \left[\left(1 - s^L\right) \mathcal{N}_t + q(\theta_t) \mathcal{V}_t \right] \tag{1.4}$$

where $\mathbb{E}_t$ is the expectation operator, M_{t+1} is the representative household's stochastic discount factor between periods t and $t+1$, and (1.3) and (1.4) are the laws of motion for open job vacancies and employment, respectively.

We assume in equation (1.3) that a new project matched with a creditor becomes an open vacancy and begins the recruiting process within the period. These vacancies join the pool of vacant positions that did not match in the previous period, $(1 - q(\theta_{t-1}))\,\mathcal{V}_{t-1}$, and those position that lost their worker, $s^L\mathcal{N}_{t-1}$, as long as the position was not also hit by a credit match termination shock s^C. Equation (1.4) assumes that a successful meeting between a firm and worker begins production the following period, again, as long as the position is not hit by a credit match termination shock s^C between the time of meeting and the start of the following period.

The asset values of a project in the three stages described above - search in the financial market, search in the labor market, and production - are found by differentiating the firm's value function. Denote these marginal asset values by $S_{j,t}$ with $j = c, l$ or g, standing for, respectively, the credit, labor and goods markets, corresponding to the market in which a project is currently operating. We have:

$$
\begin{aligned}
S_{c,t} &= -\kappa_I + p_t S_{l,t} + (1 - p_t)\mathbb{E}_t M_{t+1} S_{c,t+1}, && (1.5)\\
S_{l,t} &= \left(1 - s^C\right)\mathbb{E}_t M_{t+1}\left[q_t S_{g,t+1} + (1 - q_t) S_{l,t+1}\right] && (1.6)\\
&\quad + s^C \mathbb{E}_t M_{t+1}\left[S_{c,t+1}\right], \\
S_{g,t} &= X_t - W_t - \Psi_t + \left(1 - s^C\right)\mathbb{E}_t M_{t+1}\left[\left(1 - s^L\right) S_{g,t+1} + s^L S_{l,t+1}\right] && (1.7)\\
&\quad + s^C \mathbb{E}_t M_{t+1}\left[S_{c,t+1}\right].
\end{aligned}
$$

Equation (1.5) states that, at the margin, an additional project $\mathcal{N}_c$ reduces the firm's value by the cost of search κ_I within a time period, and pays off with two possible marginal values going forward. Either search is successful, with probability p_t, in which case the effect is valued by the firm at the margin by S_{lt+1}, or it is not. Equation (1.6) states that, at the margin, a vacant job position that is not randomly separated in the credit market affects the firm's value through the possibility of matching with a worker. With probability q_t the position is filled, and has value to the firm S_{gt+1}. All filled positions, described in equation

(1.7), generate a profit flow $(X_t - W_t - \Psi_t)$, and continue into the next period as a filled position with probability $\left(1 - s^C\right)\left(1 - s^L\right)$.

1.3.3 Financial Institutions

Financial institutions provide liquidity to firms in the labor-recruiting stage. This occurs either following a new match with a project, which results in the entry of a new vacancy, or following a labor turnover shock s^L. These institutions, owned by the representative household, maximize profits by setting an amount of potential new credit relationships $\mathcal{B}_{ct}$, searching for new investment projects at individual per period cost κ_{Bt}. These search costs are subject to exogenous, stationary, shocks. As a large institution, it pays an outflow $\gamma\mathcal{V}_t$ for the recruiting activities of each vacant position $\mathcal{V}_t$, and receives payment Ψ_t from each of the $\mathcal{N}_t$ filled positions. The financial institution's decision problem is given by

$$B_t = \max_{\mathcal{B}_{ct}} \left[\Psi_t \mathcal{N}_t - \gamma \mathcal{V}_t - \kappa_{Bt}\mathcal{B}_{ct}\right] + \mathbb{E}_t M_{t+1}\left[B_{t+1}\right] \tag{1.8}$$

$$\text{subject to} \qquad \mathcal{V}_t = \left(1 - s^C\right)\left[\left(1 - q(\theta_{t-1})\right)\mathcal{V}_{t-1} + s^L\mathcal{N}_{t-1}\right] + \bar{p}(\phi_t)\mathcal{B}_{ct} \tag{1.9}$$

$$\mathcal{N}_{t+1} = \left(1 - s^C\right)\left[\left(1 - s^L\right)\mathcal{N}_t + q(\theta_t)\mathcal{V}_t\right]. \tag{1.10}$$

Equation (1.9) is equivalent to (1.3) in the firm's problem, with the flow of new matches in the financial market expressed as matched searching creditors, $\bar{p}_t\mathcal{B}_{ct}$. Likewise, the financial intermediary is subject to the law of motion for employment (1.4) which governs its revenue stream $\Psi_t\mathcal{N}_t$.

The marginal asset values of each of the three stages of a project to a financial institution, denoted by $B_{j,t}$, $j = c$, l or g respectively, are obtained from differentiation of the financial

institutions value function:

$$B_{c,t} = -\kappa_{Bt} + \bar{p}_t B_{l,t} + (1-\bar{p}_t)\mathbb{E}_t M_{t+1} B_{c,t+1}, \tag{1.11}$$

$$B_{l,t} = -\gamma + \left(1-s^C\right)\mathbb{E}_t M_{t+1}\left[q_t B_{g,t+1} + (1-q_t)B_{l,t+1}\right] + s^C \mathbb{E}_t M_{t+1} B_{ct+1}, \tag{1.12}$$

$$B_{g,t} = \Psi_t + \left(1-s^C\right)\mathbb{E}_t M_{t+1}\left[(1-s^L)B_{g,t+1} + s^L B_{l,t+1}\right] + s^C \mathbb{E}_t M_{t+1} B_{c,t+1}. \tag{1.13}$$

Adding an additional unit of search in the financial market, by equation (1.11), reduces the financial intermediary's value by flow cost κ_B. With probability $\bar{p}_t$, however, a project is found within the period adding the value B_{lt}. Being in a match with a project in the labor-market search stage is costly to the financial intermediary. It involves a per period cost γ (see equation (1.12)). Once the position is matched with a worker, which occurs with probability q_t per period, this adds to the value of the financial institution. As the last equation (1.7) states, the financial intermediary receives payments Ψ_t each period until the either the financial market match or the labor market match are destroyed.

1.3.4 Representative Household

The household is composed of a continuum of members of unit mass who are either employment or unemployed. The employed earn per period wage W_t. The unemployed have utility from leisure $l > 0$, search for a job, and receive unemployment compensation $b > 0$. Household members pool resources, and the household chooses an aggregate level of consumption C_t, over which they have preferences $u(C)$ with the usual properties, and holdings of risk free bonds A_t to maximize:

$$H_t = \max_{C_t, A_t}\left[u(C_t) + l\mathcal{U}_t\right] + \beta\mathbb{E}_t\left[H_{t+1}\right] \tag{1.14}$$

$$\text{subject to} \qquad W_t\mathcal{N}_t + b\mathcal{U}_t + A_{t-1}(1+r_{t-1}) + D_t^S + D_t^B = C_t + T_t + A_t \tag{1.15}$$

and subject to the laws of motion for employment and unemployment. The terms $D_t^S = X_t\mathcal{N}_t - W_t\mathcal{N}_t - \Psi_t\mathcal{N}_t - \kappa_I\mathcal{N}_{Ct}$ and $D_t^B = \Psi_t\mathcal{N}_t - \gamma\mathcal{V}_t - \kappa_{Bt}\mathcal{B}_{ct}$ in the budget constraint are period profits from firms and financial institutions, respectively, rebated lump sum at the end of the period.

The marginal value of an additional unemployed and employed worker, respectively, are obtained by differentiating the household's value function:

$$\begin{aligned}
\frac{H_{Ut}}{\lambda_t} &= Z_t + \beta\mathbb{E}_t\frac{\lambda_{t+1}}{\lambda_t}\left[f(\theta_t)\frac{H_{Nt+1}}{\lambda_{t+1}} + (1-f(\theta_t))\frac{H_{Ut+1}}{\lambda_{t+1}}\right], \\
\frac{H_{Nt}}{\lambda_t} &= W_t + \beta\mathbb{E}_t\frac{\lambda_{t+1}}{\lambda_t}\left[\left(1-s^C\right)\left(1-s^L\right)\frac{H_{Nt+1}}{\lambda_{t+1}} + \left(s^C + \left(1-s^C\right)s^L\right)\frac{H_{Ut+1}}{\lambda_{t+1}}\right].
\end{aligned}$$

An unemployed worker adds $Z_t = b + l/\lambda_t$ per period to the household value, where λ_t is the marginal utility of consumption C_t, and, if search is successful - with probability $f(\theta_t)$ - adds an additional employed worker to the household. The employed workers are valued for the wage earned every period, and with probability $\left(1-s^C\right)\left(1-s^L\right)$, in the subsequent period.

1.3.5 Bargaining and Equilibrium in the Financial Market

The first order conditions for the household's problem in (1.14) yield the standard Euler equation relating the risk-free rate to expected aggregate consumption growth:

$$\frac{1}{1+r_t} = \mathbb{E}_t\beta\left[\frac{u_c\left(C_{t+1}\right)}{u_c\left(C_t\right)}\right] \equiv \mathbb{E}_t M_{t+1}. \tag{1.16}$$

A firm and a financial institution's decisions in the financial market, given by their respective choices of $\mathcal{N}_{ct}$ and $\mathcal{B}_{ct}$ which solve problems (1.2) and (1.8), satisfy the following

optimality conditions

$$\frac{\kappa_I}{p(\phi_t)} = S_{lt}, \tag{1.17}$$

$$\frac{\kappa_{Bt}}{\bar{p}(\phi_t)} = B_{lt}. \tag{1.18}$$

Choices in the financial market ensure that the marginal impact on the value functions of firms and financial institutions are equal to zero. In other words, $S_{ct} = 0$ and $B_{ct} = 0$ at the optimum. This is a free entry condition in the financial market that leads to equations (1.17) and (1.18). These state that in equilibrium the value of an open job vacancy to either the firm or the financial institution is equal to the average, respective, search costs in the financial market required to form a match.

After contact, the creditor and the firm engage in bargaining to determine Ψ_t, which denotes the creditor's share of the total match surplus $(B_{l,t} + S_{l,t})$. The repayment is negotiated each period and solves the problem:

$$\mathbb{E}_t\left[\Psi_{t+1}\right] = argmax\left(B_{lt} - B_{ct}\right)^{\alpha_C}\left(S_{lt} - S_{ct}\right)^{1-\alpha_C}, \tag{1.19}$$

where $\alpha_C \in (0,1)$ is the creditor's bargaining power relative to that of the firm. With $\alpha_C = 0$ the creditor leaves all the surplus to the firm. The solution to the generalized Nash bargaining problem is an agreed to repayment rule such that the current match surplus is shared as:

$$(1-\alpha_C)B_{l,t} = \alpha_C S_{l,t}. \tag{1.20}$$

The expected repayment rule that solves this Nash bargaining problem is:

$$\mathbb{E}_t\left[\Psi_{t+1}\right] = \alpha_C \mathbb{E}_t\left[X_{t+1} - W_{t+1}\right] + (1-\alpha_C)\left[\frac{\gamma}{q_t}\left(\frac{1+r_t}{1-s^C}\right) - \left(1-s^L\right)\mathbb{E}_t\left[\frac{\gamma}{q_{t+1}}\right]\right].$$

The above expression for the negotiated repayment states that the creditor will receive a fraction α_C of the expected profit flow from labor at date $t+1$. The second term represents

how the creditor will receive more if the current costs to fill a vacancies γ/q_t - which are being paid by the creditor in the period of price setting - are large relative to what they are expected to be in the future.

Combining (1.17), (1.18) and (1.20), we obtain the equilibrium value of credit-market tightness ϕ_t:

$$\phi_t = \frac{1-\alpha_C}{\alpha_C}\frac{\kappa_{Bt}}{\kappa_I}. \tag{1.21}$$

Financial-market tightness is decreasing in the creditor's bargaining power α_C. Increasing the share of the economic rents given to the creditor of a financial market match leads to more entry of creditors relative to investment projects. Likewise, a shock increasing the cost of search for financial intermediaries κ_{Bt} will reduce entry by creditors, and increase market tightness.

1.3.6 Return on loans

The expected rate of return on loans to firms, R_t, is the rate which sets the expected discounted value of a loan, $\frac{\gamma}{R_t+q(\theta)}$ equal to the expected discounted repayment on the loan $\frac{q(\theta)}{R_t+q(\theta)}\frac{E_t[\Psi_{t+1}]}{R_t+s^C+(1-s^C)s^L}$ (as in *Wasmer and Weil* (2004) and *Petrosky-Nadeau* (2013)). This results in a an expected return from lending in the credit market:

$$R_t \;=\; \frac{\mathbb{E}_t[\Psi_{t+1}]}{\gamma/q(\theta_t)} - \left(s^C + \left(1-s^C\right)s^L\right). \tag{1.22}$$

The expected return depends, first, on the expected flow of repayments to the creditor relative to the size of the outflow during the labor-market recruiting stage, $\gamma/q(\theta)$. The second term corresponds to discounting from the termination of the repayment flow. An increase in α_C results, all else equal, results in a greater flow repayment Ψ_t.

1.3.7 Equilibrium in the Labor Market

The total amount of search costs in financial markets involved in creating a position in a firm, those associated with creating a new financial relationship, are summarized by the variable:

$$K_t \equiv \frac{\kappa_I}{p(\phi_t)} + \frac{\kappa_{Bt}}{\bar{p}(\phi_t)}. \tag{1.23}$$

These costs represent the value of a match in the financial market to both parties, or their joint surplus $(B_{l,t} + E_{l,t})$. The marginal values from a creditor-project match in the labor recruiting stage l and the production stage g are given by

$$S_{lt} + B_{l,t} \equiv F_{lt} = -\gamma + \left(1 - s^C\right) \mathbb{E}_t M_{t+1} \left[q_t F_{g,t+1} + (1 - q_t) F_{l,t+1}\right] \tag{1.24}$$

$$S_{gt} + B_{g,t} \equiv F_{gt} = X_t - W_t + \left(1 - s^C\right) \mathbb{E}_t M_{t+1} \left[(1 - s^L) F_{g,t+1} + s^L F_{l,t+1}\right]. \tag{1.25}$$

At any date the value of a vacant position in the labor market to the creditor-project pair is equal to the current value of its creation costs in the financial market, K_t. Equation (1.25) is reminiscent of the expression for the value of a job vacancy in the standard DMP model, and converges to such an expression when $s^C = 0$. A free-entry equilibrium without search frictions would have the value of F_{lt} converge to 0 at all dates.

Substituting $F_{lt} = K_t$ in equation (1.24) we have

$$\frac{K_t + \gamma}{q(\theta_t)} = \left(1 - s^C\right) \mathbb{E}_t M_{t+1} \left[F_{gt+1} + \left(\frac{1 - q(\theta_t)}{q(\theta_t)}\right) K_{t+1}\right]. \tag{1.26}$$

This job-creation condition equates the expected costs from financial-market and labor-market search to the expected benefit from filling the position (conditional on the financial-market match surviving to the next period). In the limit, as K_t tends to zero at all dates we recover the canonical DMP job-creation condition. In the presence of a frictional financial market, the right-hand side has an additional term $\left(1 - s^C\right)(1 - q_t)\,\mathbb{E}_t M_{t+1} K_{t+1}/q_t$. This

captures the value of an unfilled vacancy in the event search is not successful in the period, and the position survives into the next.

By defining a variable to summarize the job-creation costs net of the position's value in the event of unsuccessful search as $\Gamma_t = \frac{K_t+\gamma}{(1-s^C)} - (1-q_t)\,\mathbb{E}_t M_{t+1} K_{t+1}$, we obtain the job creation condition for the model with search frictional credit and labor markets

$$\frac{\Gamma_t}{q_t} = \mathbb{E}_t M_{t+1}\left[X_{t+1} - W_{t+1} + \left(1-s^C\right)\left[\left(1-s^L\right)\frac{\Gamma_{t+1}}{q_{t+1}} + s^L K_{t+1}\right]\right]. : \tag{1.27}$$

The wage is the solution to a Nash bargaining problem between the worker and the firm. It is the solution to the problem:

$$W_t = argmax\left(\frac{H_{Nt} - H_{Ut}}{\lambda_t}\right)^{\alpha_L}\left(F_{gt} - F_{lt}\right)^{1-\alpha_L}. \tag{1.28}$$

The worker-firm negotiated wage must satisfy the usual sharing rule

$$\alpha_L\left(F_{gt} - K_t\right) = \left(1-\alpha_L\right)\left(H_{Nt} - H_{Ut}\right)/\lambda_t,$$

and the resulting wage is:

$$\begin{aligned} W_t &= \alpha_L\left(X_t + \theta_t\left[\frac{\gamma}{(1-s^C)} + \left[\frac{r_t + s^C}{(1-s^C)(1+r_t)}\right]K_t\right]\right) \\ &\quad + \left(1-\alpha_L\right)Z_t - \alpha_L\left[\frac{r_t+s^C}{1+r_t}\right]K_t. \end{aligned} \tag{1.29}$$

1.3.8 Stochastic processes

Labor productivity and the cost of search for financial institutions follow stationary AR(1) processes in logs. That is, we have $\log X_t = \rho_x \log X_{t-1} + \nu_{xt}$, where $0 < \rho_x < 1$ and ν_{xt} is white noise for labor productivity. In the financial market, the search costs are assumed to follow $\log \kappa_{Bt} = (1-\rho_{\kappa_B})\log \overline{\kappa}_B + \rho_{\kappa_B}\log \kappa_{Bt-1} + \nu_{\kappa_B t}$. The innovations ν_{xt} and $\nu_{\kappa_B t}$ are assumed to be independent. It is certainly possible that credit and productivity shocks are

correlated in reality, and we aim to explore this concept in future work.

1.3.9 Equilibrium

The household's budget constraint leads to the economy's aggregate resource constraint

$$Y_t \;=\; C_t + \gamma V_t + \kappa_{Bt}\mathcal{B}_{ct} + \kappa_I \mathcal{N}_{ct}. \tag{1.30}$$

An equilibrium is defined as a set of functions from labor market tightness θ_t, wage W_t, credit market tightness ϕ_t, aggregate consumption C_t, risk free rate r_t, that, for current employment $\mathcal{N}_t$, productivity X_t, credit market search costs κ_{Bt}, solve: (i) the job creation condition (1.27); (ii) Nash wage rule (1.29); (iii) credit market equilibrium condition (1.21); (iv) consumption Euler equation (1.16), and; (v) aggregate resource constraint (1.30).

1.4 Quantitative Results

The model is calibrated to match key features of U.S. labor and credit markets. Our calibration strategy is described in Section 1.4.1 and the results are given in Section 1.4.2. Section 1.4.3 assesses the effects of productivity and credit-market shocks on labor market, and examines how those effects vary across recession, normal, and expansion periods as defined in our empirical exercise in Section 1.2.

1.4.1 Parameterization and calibration

Our model has many parameters. We use commonly accepted values where appropiate, but otherwise resist the temptation to use parameter values from the labor market search literature. Most models in this literature do not have search-frictional credit markets, and borrowing parameters from this literature would effectively relegate credit markets to an appendage of a pre-calibrated labor-market model. Instead, we take an approach that provides credit and labor markets frictions an equal chance at influencing the data generating process.

We do this by performing a comprehensive search over a discretized parameter space and examining the model's key moments at each parameterization. While this approach is more cumbersome than a traditional calibration exercise, it prevents us from setting parameters to values that are sensible in one family of models but not necessarily this family of models. The outcome of this exercise is summarized in Table 1.2 and the corresponding moments generated by the data are discussed in Section 1.4.2.

Table 1.2: Model Parameters

	Parameter	Value
Labor market:		
job-separation rate	s^L	0.035
worker bargaining weight	α_L	0.40
vacancy cost	γ	0.10
labor matching curvature	η_L	1.15
non-employment value	z	0.84
Credit market:		
credit-match separation rate	s^C	0.01/3
creditor bargaining weight	α_C	0.30
mean search costs	$\overline{\kappa}_B = \kappa_I$	0.16
credit matching curvature	η_C	1.2
discount factor	β	.997
Shock processes:		
productivity persistence	ρ_x	$0.95^{1/3}$
productivity volatility	σ_x	0.0065
credit search persistence	ρ_{κ_B}	$0.95^{1/3}$
credit search volatility	σ_{κ_B}	0.0065

Credit market parameters: β, s^C, η_C, α_C, κ_I, and $\overline{\kappa}_B$

The discount factor β is set such that the risk-free rate r averages an annualized 4%. The separation rate s^C is set for a 1% quarterly firm exit rate. The credit matching function $M_c(\mathcal{B}_{ct}, \mathcal{N}_{ct}) = \mathcal{N}_c\mathcal{B}_c / \left(\mathcal{N}_c^{\eta_C} \mathcal{B}_c^{\eta_C}\right)^{1/\eta_C}$ as in *Den Haan et al.* (2000), and set the curvature parameter $\eta_C = 1.2$. This value is comparable to the curvature parameter used in a labor-

market search model in *Den Haan et al.* (2000), which is 1.27. The creditor's bargaining weight resulting from our parameter search is $\alpha_C = 0.30$ which is notably smaller than the values between .5 and .9 seen in *Petrosky-Nadeau and Wasmer* (2012). To keep our analysis straightforward, we set $\kappa_I = \bar{\kappa}_B$ and find the best performance when $\kappa_I = .16$. This is considerably larger than the cost of credit-market search in *Petrosky-Nadeau and Wasmer* (2012), though our model differs in several ways.[4]

Labor market parameters: s^L, η_L, γ, z, and α_L

The rate of labor separation s^L is set to a monthly rate of 0.035, consistent with the estimate based on JOLTS data and accounting for concurrent separations from credit-market separations. The curvature parameter in the matching function $\mathcal{N}_l\mathcal{U}/\left(\mathcal{N}_l^{\eta_L} + \mathcal{U}^{\eta_L}\right)^{\frac{1}{\eta_L}}$, is set to $\eta_L = 1.15$. As was the case with η_C, this is well below values seen in comparable models without credit-market frictions. This suggests that the presence of additional frictions allows matching functions to be less elastic in order for their combined effect to compare to a model with only one friction. The parameter search process yields a labor-market search cost of $\gamma = 0.10$, a value of unemployment of $z = .84$, and a worker's share of wage bargaining of $\alpha_L = .40$, which is larger than values in *Petrosky-Nadeau and Wasmer* (2012) which range from .03 to .15. The cost of searching in the labor market, γ, is comparable to *Petrosky-Nadeau and Wasmer* (2012), while the our value of unemployment is notably higher as *Petrosky-Nadeau and Wasmer* (2012) uses values between .4 and .7. We are, however, well below other values seen in the labor market literature such as in *Hagedorn and Manovskii* (2008) where z would be closer to .9.

Shock process parameters: ρ_x, σ_x, ρ_{κ_B}, σ_{κ_B}.

The basic unit of time is a month. The process for productivity described in Section 1.3.8 has a persistence parameter $\rho_x = 0.95^{1/3}$, and conditional volatility, $\sigma_x = 0.0065$. These are standard parameter values in line with the volatility and persistence of labor productivity

[4]Key differences include the introducing stochastic search costs, the use of a *Den Haan et al.* (2000) matching function as opposed to a Cobb-Douglas matching function, and exogenous separation of firm-creditor relationships.

measure by the BLS. For symmetry, we use matching values for ρ_{κ_B} and σ_{κ_B} and allow the credit and labor market parameters to bring model volatility to values comparable to the data.

1.4.2 Stationary and business cycle moments

The "best" parameters are chosen by comparing the following moments between the model and the data. The observed credit spread is simply the BAA - 10 year treasury spread. The credit spread in the model is computed by comparing the return on loans with and without credit frictions. The special case of competitive prices in the financial market with no surplus to the creditor arises when $\alpha_C = 0$. The returns on loans in this competitive pricing are denoted R_t^*. The model's friction credit spread is simply the difference between the bargained R_t and the competitive R_t^*.

The moments of interest include first moments and autocorrelations of unemployment and the credit spread observed across all periods. Additionally, we consider second moments of unemployment and the credit spread as observed in different economic states. Our interest in state-dependence arises from our empirical findings in Section 1.2.2 where we observed a larger response in unemployment following credit-spread increases during recessions. We consider two economic states, recession and normal, which adhere to the same definitions used in our empirical analysis (see Section 1.2.1). In addition to these first and second moments, we examine key correlations between productivity, vacancies, unemployment, and the credit spread in normal and recession states. The moments for both observed and simulated data are given in Table 1.3.

The model performs well, though slightly high, on first-order moments. The model generates a mean unemployment of 6.26 compared to 5.92 in the data and a mean credit spread of 2.44 compared to 1.90 in the data. For second-order moments, the model generally matches the pattern of increased volatility during recessions and does a good job overall at matching volatility levels in both states. The model generates slightly less volatility for

Table 1.3: Moments from Observed and Simulated Data

Normal States:

	U		Spread	
	Data	Model	Data	Model
Mean	5.92	6.26	1.90	2.44
St. dev.	0.12	0.08	0.19	0.20
Autocorrelation	0.97	0.93	0.93	0.91
Correlation w. X_t	-0.10	-0.58	-0.23	-0.44
Correlation w. V_t	-0.93	-0.88	-0.59	-0.80
Correlation w. R_t	0.55	0.81	-	-

Recession States:

	U		Spread	
	Data	Model	Data	Model
St. dev.	0.16	0.16	0.26	0.33
Correlation w. X_t	-0.50	-0.28	-0.65	-0.43
Correlation w. V_t	-0.93	-0.70	-0.73	-0.81
Correlation w. R_t	0.57	0.83	-	-

Notes: Unemployment rate data obtained from the BLS for the population over 16 years of age. The credit spread is the difference between the return on BAA corporate bonds and the 10 year Treasury notes. Model moments are from 5000 simulations of 2,232 months each, transforming to quarterly values by simple averaging. Standard deviations and correlations are computed for HP-filtered deviations from means.

unemployment during normal periods with a standard deviation of 0.08 compared to 0.12 in the data. The volatility of the simulated credit-market spread, on the other hand, closely matches the data during normal times, with respective standard deviations of 0.19 and .20. During recessions, however, it is unemployment that performs best in terms of matching the volatility observed in the data. Unemployment in both the model and data has standard deviations of 0.16 during recessions. The credit spread in the model demonstrates increased volatility during recessions as does the data, with a standard deviation of 0.33 compared to 0.26 in the data.

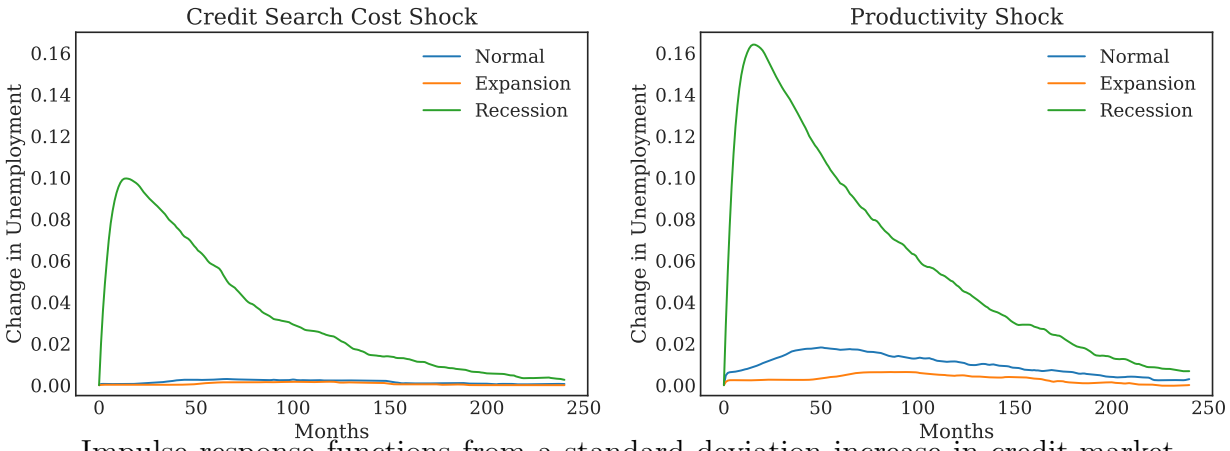
For correlations, the model does well at matching the signs and relative magnitudes observed in the data. In both the model and data, the correlation between productivity and unemployment is negative as is the correlation between unemployment and vacancies. The strength of the correlation with unemployment is stronger for vacancies than productivity, which is true for both the model and the data. Key correlations include the correlation between unemployment and vacancies (the Bevridge curve), which is as measured by as -0.88 in the model and -0.93 in the data. Another key correlation is the relationship between unemployment and the credit spread, which is 0.83 in the data and 0.57 in the data. Neither of these key correlations differ between normal and recession states. Other correlations, such as the relationship between productivity and unemployment, become stronger in the data during recession states while they become weaker in our model simulated data. This is a point that future research should aim to investigate and improve upon.

1.4.3 State Dependance and the transmission of shocks

We compute impulse response functions to measure the impact of adverse productivity and financial-market shocks on unemployment in our model. To examine state dependence, we consider different initial conditions for each impulse response function. Our three key state variables are the employment rate, productivity, and the cost of credit-market search, and we consider initial conditions spanning the distributions of these variables. During recession

periods, mean unemployment is 10% and our productivity and credit-market variables are near their 5th and 95th percentiles, respectively. In normal periods, mean unemployment is 5% and our productivity and credit-market variables are near their median values. When the economy is in an expansion period, mean unemployment is 4%, productivity is near its 5th percentile and the credit-market is near its 95th percentile.

Figure 1.5: Impulse responses for Unemployment



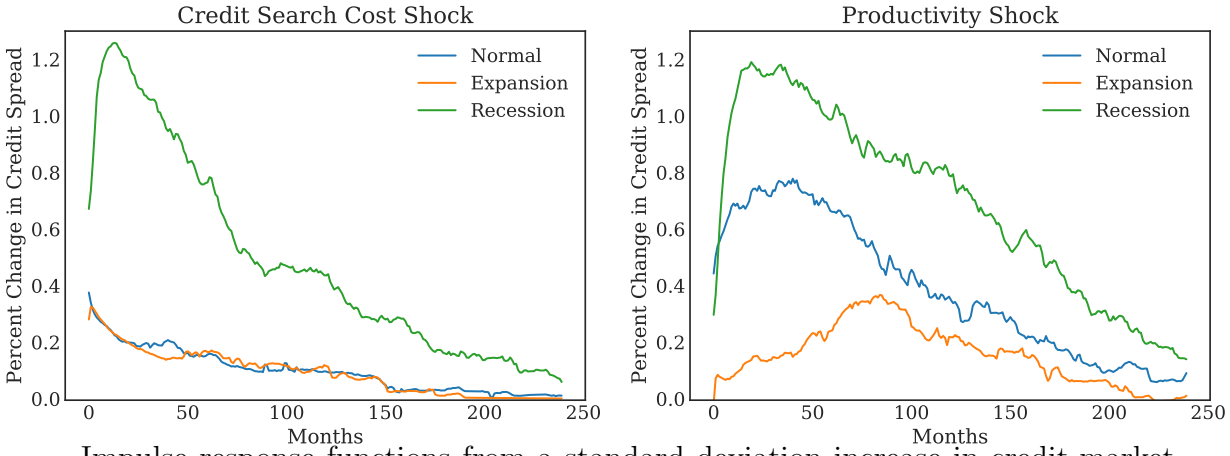
Impulse response functions from a standard deviation increase in credit-market search cost (left), and a standard deviation decrease in productivity (right). IRFs are computed as the mean of 500 simulations.

Our results are shown in Figure 1.5. Just as in the search-frictional models in *Petrosky-Nadeau and Zhang* (2013b), *Petrosky-Nadeau and Wasmer* (2012) and *Petrosky-Nadeau* (2013), the impulse response functions differ dramatically with the initial state of the economy. Due to the nonlinearities nature of the matching functions, and because we solved our model relying on linear approximations, unemployment is much more sensitive to shocks when the labor market is already slack. In our model a standard deviation increase in the cost of credit-market search during a recession leads to an increase in unemployment of nearly 10 percentage points. The same shock during an expansion or normal state has no noticeable affect on unemployment. The same is true of productivity shocks. A standard deviation decrease in productivity during a recession results in a 16 percentage point increase in unemployment, while the same shock during normal states results in only a 2 percentage

point increase in unemployment. In both credit-search and productivity shocks, the response in unemployment peaks at around 20 months after the shock.

Note that the impulse response functions are not immediately comparable to the empirical exercise in Section 1.2. Those results examine the relationship between the unemployment and the credit spread itself, whereas these results examine the relationship between unemployment and credit-market search. The general pattern of state dependence, however, is present in both settings and we are encouraged that our structural model and unique calibration strategy yield a model that demonstrates state dependence as we found empirically.

Figure 1.6: Impulse responses for Credit Spread



Impulse response functions from a standard deviation increase in credit-market search cost (left), and a standard deviation decrease in productivity (right). IRFs are computed as the mean of 500 simulations.

The credit spread also demonstrates state dependence in its response to both productivity and credit-market search costs. Figure 1.6 shows the impulse response functions for the credit market spread in response to a standard deviation increase in the cost of credit market search (left panel), and a standard deviation decrease in productivity (right panel). During normal and expansion states, a standard deviation increase in the cost of credit search results in a 40% increase in the credit spread. During recessions, however, the response is over three times larger with more than a 125% increase. The response resulting from to productivity shocks are similar to those caused by credit shocks, and differ mostly in that productivity shocks

make a larger impact during normal and expansion periods than shocks to the cost of credit-market search. A standard deviation decrease in productivity increases the credit spread in recession, normal, and expansion states by, respectively, 120%, 75%, and 35%. While unemployment is about 1.6 times more responsive to productivity shocks during recessions, the response of the credit spread is about even between productivity and credit-market shocks. This insight is new to the search and matching framework as we are the first to use stochastic credit-market shocks which allows us to compute their resulting impulse response functions.

1.5 Conclusion

The relationship between credit and labor markets will continue to be an important field of study in future decades. In this study we provide empirical evidence that the relationship between these two markets demonstrates a marked degree of state dependence; the impact of an adverse shock will be much larger during recessions than at other stages of the business cycle. We also make theoretical contributions in the form of a model with search-frictional credit and labor markets, which we bring to the data through a rigorous calibration exercise. The model matches all of our major labor market moments, including a matrix of correlation coefficients. The moments are examined and matched for both normal times and periods of economic recession. The impulse response functions demonstrate the same kind of state dependence as seen in our empirical exercise, and suggest that credit market fluctuations can have potentially large impacts if the economy is already in a poor state. Both credit-market and productivity shocks are able to generate large movements in unemployment and the credit spread, but unemployment is about 1.6 times more responsive to productivity shocks than credit shocks when the economy is in a poor initial state.

There are many ways to extend and improve upon our current analysis. One could use this model to better understand the most likely sequence of shocks to match an event such as the 2008 recession or other historical events. This kind of analysis would provide insight into

the relative importance, in a search-frictional framework, of productivity shocks as compared to financial market shocks. Mapping this model to the data would also be improved by the discovery of any empirical counterpart to κ_b, the cost of searching within the credit-market. This would enable a true mapping of model fundamentals to the data as opposed to our current approach of mapping model outputs to data. Future studies might examine credit constraints as they relate to capital or other expenses which are more likely to require outside financing, as opposed to hiring costs which are relatively inexpensive for many firms. Future work might also consider the impact of monetary policy interventions in search-frictional markets, and how the impact of those interventions vary over the business cycle.

CHAPTER II

Market Expansion and Market Concentration

2.1 Introduction

For most of human history, geographic distance has played an important role in determining the extent to which different firms act as competitors. When two firms are sufficiently distant, each can make decisions without fear of losing customers to the other. Over the past century, however, the concept of distance has changed in ways that have changed markets in fundamental ways. Across several industries, changes in both technology and policy have steadily removed barriers to trade, geographic or otherwise, such that firms can reach customers located farther and farther away. These changes have dramatically changed the economic environment for many affected industries as market expansion has pitted firms against one another across larger geographies.

The pattern of market expansion leading to higher market concentration is evident in several industries. Well studied examples include the US commercial banking industry, the US brewing industry, international trade and manufacturing, and US retail. Studies for each of these examples, however, largely appear in isolated literatures. The siloed nature of these literatures makes it difficult to recognize that even though these industries differ in many ways, all of these examples follow the same pattern: markets expansion leads to a reduction in the number of firms and increased market concentration. Motivated by this pattern, I present a novel theoretical model featuring heterogeneous firms competing within a single industry

but in different locations a la *Hotelling* (1929). Firms compete more intensely against firms located in close proximity and do not compete at all against firms located beyond a given radius. Firms may endogenously exit if their idiosyncratic and persistent shocks become unfavorable. Within this theoretical framework, I examine the impact of market expansion by increasing the radius at which firms are unable to compete, and compare the results of both sudden and gradual changes to this parameter. The model succeeds in reproducing the aggregate behavior of firms in the face of market expansion as observed in various industries. Market expansion places downward pressure on prices through increased competition. Firm exit, however, places upward pressure on prices in the short-run by shifting out the demand curves among surviving firms. Whether prices increase or decrease in the long run depends only on the size of the market expansion and not the speed at which the expansion occurs.

In addition to juxtaposing these literatures in a way that makes this pattern apparent, in this paper I present a theoretical model with heterogeneous firms distributed over an abstract geographic space. The model explicitly parameterizes the geographic size of the market, allowing firms to compete across larger distances over time. The model is general enough to include, as special cases, an economy comprised of isolated monopolists and also a dynamic oligopoly model with endogenous entry and exit. The model successfully duplicates the key pattern of market expansion driving market concentration, and allows me to compare outcomes between sudden and gradual episodes of market expansion.

The novel feature in the model is an inverse demand curve that explicitly accounts for the geographic distance between a firm and its competitors. The price for a given firm is a decreasing function of its own output as well as the output of all competing firms *w*eighted by distance. The weight is only positive within a specified radius, such that a firm will not be influenced by a competing firm unless it is within the given radius. Distance is achieved by assigning each firm a set of coordinates in a compact metric space. Weights are given by a function that resembles a kernel density estimator, where the radius of interest is determined by the kernel's bandwidth parameter. By increasing the bandwidth parameter,

firms are subject to competition from firms that were formerly too distant, as is the case when markets expand due to changes in policy or technology.

A key assumption of the model is that firms do not attempt to forecast the states of relevant competitors as they would under rational expectations. Rather, they only keep track of the weighted sum of competitor output. While this reduces the dimensionality of the firm's problem, it complicates the firm's ability to form expectations as the future weighted sum of competitor output does not evolve in a predictable way due to firm exit and random productivity shocks. Consequently, firms form naïve expectations over observable aggregate variables such as the sum of competitor output weighted by distance, and the total number of active firms. Firms observe the current aggregates and assuming they will be the same the following period. This assumption, while severe, is justified for several reasons. First, it makes the model solvable for any number of firms. Second, it simplifies the equilibrium concept to that of a simple recursive equilibrium and avoids the multiplicity of equilibria present in repeated games.

The model demonstrates rich dynamics in the face of expanding markets. Simulations begin from a point in which markets are so small each firm acts effectively as a monopolist in their own isolated economy. Over time, the parameter for market size increases and the economy evolves in response to the changing environment. In the first and simplest version of the model, exiting firms do not acquire market share from exiting firms. In this case, when markets expand, increased competition places downward pressure on prices and many firms exit as a result. Firm exit, however, relieves much of these competitive pressure and pushes prices back toward their starting values. In the long run, in spite of a smaller number of firms, the expanded markets remain more competitive and prices decline in the long run. The severity of price fluctuations varies with the speed of market expansion, which larger fluctuations occurring during rapid episodes of market expansion. In the long run, prices converge to a single, lower price regardless of market size.

Each firm's inverse demand curve shifts outward when other firms exit such that aggre-

gate demand remains unchanged. This creates another channel by which prices experience upward pressure in the short run, as demand increases for output from surviving firms while production takes time to adjust. This additional channel causes prices to fluctuate more in the short run when markets expand quickly. The long-run impact on prices varies with the degree of market expansion. The impact on market concentration, however, is nonlinear. Market expansion leads to increased firm exit which increases market concentration, but market expansion also leads to decreased output per firm which in my model prevents large firms from grabbing additional market share. The resulting levels of market concentration, however, are still much larger than in any isolated monopolist setting.

The paper continues as follows. In Section 2.2, I review the literature immediately applicable to geographic expansion and its role in enabling increased market concentration. In Section 2.3, I review the literatures of several well-studied episodes where market expansion preceded dramatic increases in market concentration. In Section 2.4, I outline the theoretical model and in Section 2.5, I describe my solution method and quantitative exercises. Section 2.7, concludes. I provide additional background information in the Appendix, including a step-by-step outline of my solution algorithm.

2.2 Literature Review

In this section, I examine studies which analyze the broader trend of market concentration across industries. The literature focused on concentration as a whole is relatively small compared to many of the literatures focused on concentration for single industries. I examine several of these industries and their respective literatures in Section 2.3.

The majority of market concentration studies, whether micro or macro, measure market concentration by the ratio of output between the four largest firms in an industry relative to the total. Colloquially, this term is known as the "four-firm industry ratio" or CR4. Some studies, where data is sufficiently detailed, include HHI or Gini values as measures of market concentration.

Mueller and Hamm (1974) use US Census data to aggregate market concentration trends across industries. They find both output-weighted and unweighted CR4 measures increase by about 2 percentage points between 1947 and 1970. To my knowledge, there are no additional studies that have brought the analysis *Mueller and Hamm* (1974) into more recent decades.

Several studies look at the relationship between market concentration and broad economic outcomes such as economic growth or innovation. *Pagano and Schivardi* (2003) find a positive correlation between economic growth and market concentration. *Beck et al.* (2008) shows financial developments interact with the distribution of firm size to disproportionately accelerate the growth of smaller firms. *Acs et al.* (1999) measures market share by the both the proportion of total employment hired by a firm and the proportion of establishments owned by a firm. They find, under either measure, that industries in which larger firms have larger market share have greater productivity growth.

One economic outcome of particular interest to many market concentration studies is the amount of R&D spending. Many speculate that larger firms would be more likely to undertake R&D projects, and so an economy with fewer and larger firms could be beneficial for innovation (*Cockburn and Henderson*, 2001). Conversely, *Mukhopadhyay* (1985) finds that R&D intense industries are correlated with *d*eclining levels of market concentration. *Bos et al.* (2013), on the other hand, suggests the relationship between market concentration and R&D follows an inverted-U shape with extremely-low and extremely-high levels of market concentration associated with low levels of investment in R&D. *Symeonidis* (1996) claims that R&D intensity and market structure are jointly determined by technology, the economic environment, and chance. None of these studies, however, are interested as much in the origins of market concentration as in the potential consequences. While discussion around the causes of market concentration are present in many industry-specific literatures (see Section 2.3), my paper uniquely brings these examples together using a common element of distance.

This paper also is relevant to the literature on geospatial competition. Paramount within

this literature is the framework proposed by *Hotelling* (1929) in which firms are mapped to a point along an interval and a customer pays a travel cost in addition to the price of a good. Where a firm is positioned, relative to both customers and other firms, determines the market clearing price. There are numerous extensions of this framework, including locations with heterogeneous products (*Phlips and Thisse*, 1982; *Irmen et al.*, 1998), strategic entry (*Lancaster*, 1982), and search-frictional goods markets (*Stahl*, 1982). The Hotelling framework has been used to explain, among other things, the trade-off between product similarity and product accessibility, why establishments selling similar products occasionally locate in close proximity (*Stahl*, 1982) – or far apart *Brown* (1989) – and the existence of price discrimination (*Norman and Nichols*, 1982). An important discussion within this literature is the similarity between geographic distance and product differentiation.[1] Both geographic distance and product differentiation serve the purpose of minimizing direct competition from neighboring establishments. Firms can set prices knowing that customers will have to engage in costly travel and search to consider products from other establishments. Beyond Hotelling, there are geospatial frameworks based on the work of Cournot competition (*Anderson and Neven*, 1991), network structures (*Economides*, 1996), and labor market search and matching (*Wasmer and Zenou*, 2006). My work differs in its emphasis on how the concept of distance has changed over time for different industries. To focus on this issue directly, I do not make location a choice as is typical in Hotelling-type models. I also limit my understanding of "distance" to a single dimension. These simplifying assumptions allow me to use the most important concepts from the Hotelling literature in an environment that is slowly but steadily changing over time.

Empirically, the literature surrounding geographic distance as it pertains to economic phenomena is surprisingly small. *Hanson* (2005) estimates a *Krugman* (1991) model using US county-level data for wages, consumption, housing stock, and exports. He finds that demand linkages between distant counties are growing between 1970 and 1990 (i.e. markets

[1] As examples, see *Stahl* (1982), *Capozza and Van Order* (1982), *Economides et al.* (1986), and *Chang et al.* (1991).

are expanding). Distance still takes a toll economic relationships, however, as counties over 1000 kilometers away demonstrate no significant economic linkages. *Evans and Harrigan* (2005) finds that geographic distance determines the location of imports in a way that reduces shipping times and inventories. *Kvasnička et al.* (2018) finds gasoline prices are lower when stations are located close to each other, suggesting that even though the commodity is essentially the same across locations, the market is local rather than national.

A separate literature, pertinent to this study, is focused on the narrower question of how to measure economic distance. It is clear from many studies that geography is only part of what comprises distance between firms and customers (or firms and other firms). Infrastructure, for example, may reduce the impact of geographic distance on economic activity between two regions. *Conley and Ligon* (2002) combine shipping rates with airline fares to estimate an infrastructure-weighted distance between countries. They use the resulting cross-section of distance to explain patterns in economic growth rates. *Tsang and Yip* (2007) measure economic distance as the difference in real GDP per capita between two countries, suggesting that developed countries are more distant to undeveloped than to other developed countries. They find that FDI hazard rates are lower in more distant countries, whether more or less developed than the host country. *Yitzhaki* (1994) uses the amount of overlap areas in two neighboring distributions to infer a measure of economic distance between industries with varying degrees of similarity.

Alternative theories as to the fundamental cause of market concentration certainly exist. *Lucas* (1978) suggests that firm size simply mirrors the distribution of managerial talent. *Luttmer* (2007) points to a process where successful firms invest in productivity improvements and smaller/newer firms imitate the larger firms. He concludes that, in the context of his model, the US distribution of firm size is the result of costly firm entrance costs... *Coad* (2010) finds a model of random growth, when combined with an accurate distribution of firm age, effectively matches the observed distribution of firm size. These studies differ from mine in that my interest is dynamic in nature. Rather than explaining the market funda-

mentals that might create a particular distribution of firm size, I am interested in how those market fundamentals are changing over time in a way that describes the observed trends in market concentration over time. Technology certainly has played a role in facilitating market expansion for some industries. *Sinai and Waldfogel* (2004) examines the extent to which the internet removes geographic barriers. *Goldmanis et al.* (2010) study a number of niche retail industries and show that the advent of e-commerce within these industries affects the distribution of market shares. In every case, a consequence of the technologically-driven market expansion is a production sector that has fewer, but larger firms.

2.3 Examples of Market Size and Market Concentration

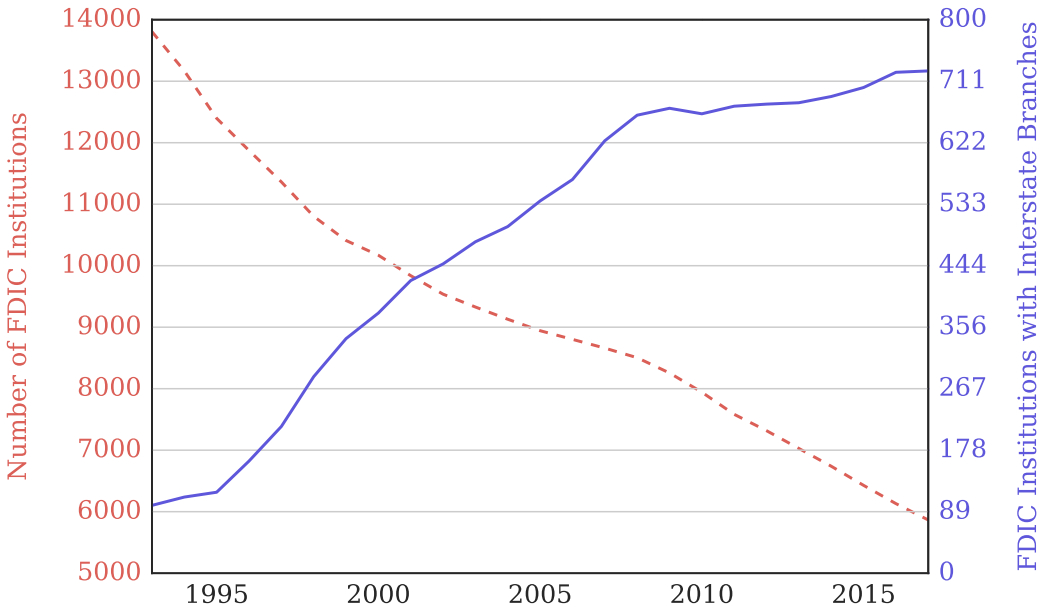
While my study is not focused on a particular industry, there are several industry-specific examples of market expansion leading to higher levels of market concentration. Each of these examples have their own literatures (with larger literatures when data is publicly available). I make no claim to cover all known episodes of industries that have been transformed by expanding markets, but include a sufficiently diverse set of examples to make a convincing argument that this phenomena is not limited to particular industries. I include examples from several US industries where markets have expanded dramatically by a combination of policy and technology including commercial banking, retail, brewing, and media/publishing industries. I also examine international markets in the wake of trade liberalization.

2.3.1 US Banking Deregulation

One striking example of market expansion is the US commercial banking industry. Prior to the 1994 Riegle-Neal Interstate Banking and Branching Efficiency Act (IBBEA), state-chartered banks were not able to establish out-of-state branches except for in states which specifically allowed it. Throughout the 1970's and 1980's, these barriers to interstate branching were removed by individual states until 1994 when the IBBEA removed all remaining barriers. As these barriers were removed, local banks suddenly faced competition from out-

of-state banks.[2] Figure 2.1 shows that in 1993, about 90 of 14,000 FDIC insured institutions operated interstate branches. In the years following, banks have exited rapidly while the number of banks with interstate branches has dramatically increased. In 2017, 711 of only 6,000 FDIC banks had interstate branches.

Figure 2.1: FDIC Institutions and Interstate Branching over Time



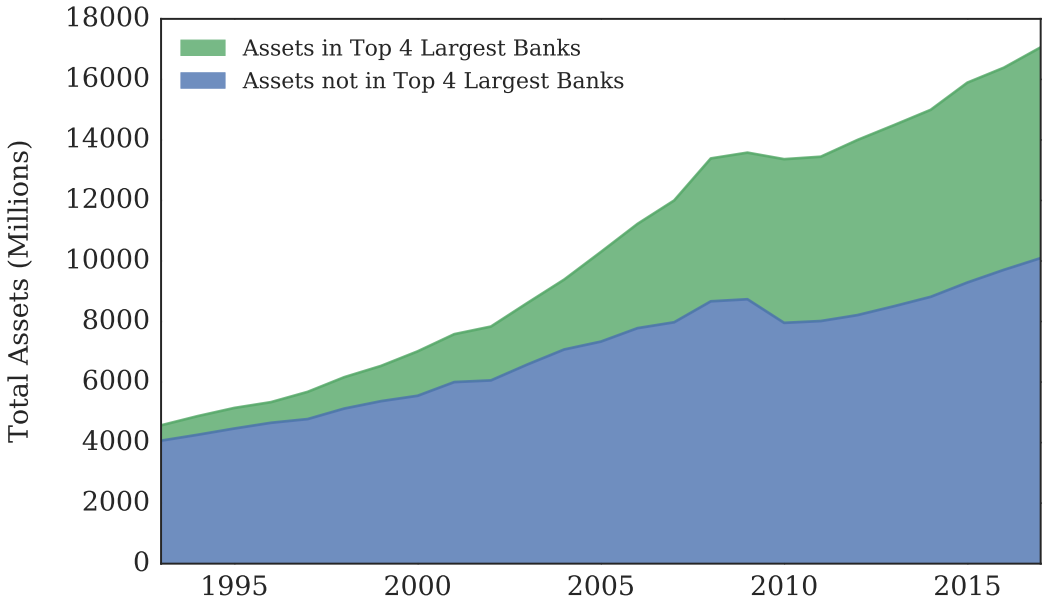
Source: author's calculations from FDIC Statistics on Depository Institutions data, Q1 values only. The red (dashed) line shows the number of FDIC insured institutions over time. The blue (solid) line shows the number of FDIC insured institutions with branches in multiple states over time.

The rapid exit of banks should not be confused with a lack of demand for banking services. Figure 2.2 shows the volume of assets held by the 4 largest banks (blue) and all other banks (green). Assets steadily increase from 1993 on, but have become increasingly concentrated among a small number of large banks. In 1993, the largest 4 banks hold about 11% of total assets. In 2017, the largest 4 banks hold about 40%.

Other studies provide further evidence of an increased market. *Petersen and Rajan* (2002) show using data from the National Survey of Small Business Finance, firms with

[2]For a history of the regulations on interstate banking, see *Kroszner and Strahan* (1999). For state-level details on deregulation process, see *Strahan* (2003). For a general survey of banking deregulation, see *Levine* (2005).

Figure 2.2: Asset Share of the Four Largest Commercial Banks over Time



Source: author's calculations from FDIC Statistics on Depository Institutions data, Q1 values only. The green region shows the volume of total assets owned by the four largest FDIC insured institutions, as measured by total assets. The blue region shows the volume of assets owned by all other FDIC insured institutions.

credit relationships established between 1973 and 1979 are located an average of 51 miles from their creditor. By comparison, firms entering into credit agreements between 1990 and 1993 did so with creditors located an average distance of 161 miles away. "While our evidence indicates that small businesses continue to use their local banks for deposit transactions, *the effective size of the credit markets faced by small firms is continuously expanding*" (pages 3-4, emphasis added)."

While this market expansion occurred while states were removing barriers to interstate banking, policy was certainly not the only driver. Many studies, including *Senzel* (1992), *Rollinger* (1996) and *Kroszner and Strahan* (1999), show improvements in technology to be key drivers in this market expansion, and facilitated the ensuing changes in policy. Technologies including the Automatic Teller Machine (ATM), electronic transfers, and the availability of banking services via phone and mail-order removed or diminished many of the geographic barriers that were protected by regulations on interstate branching. More recently, the ad-

vent of online banking has continued to redefine "local" as a requirement for consumers choosing among different bank service providers (See *Lambrecht et al.* (2006), *Allen et al.* (2008)).

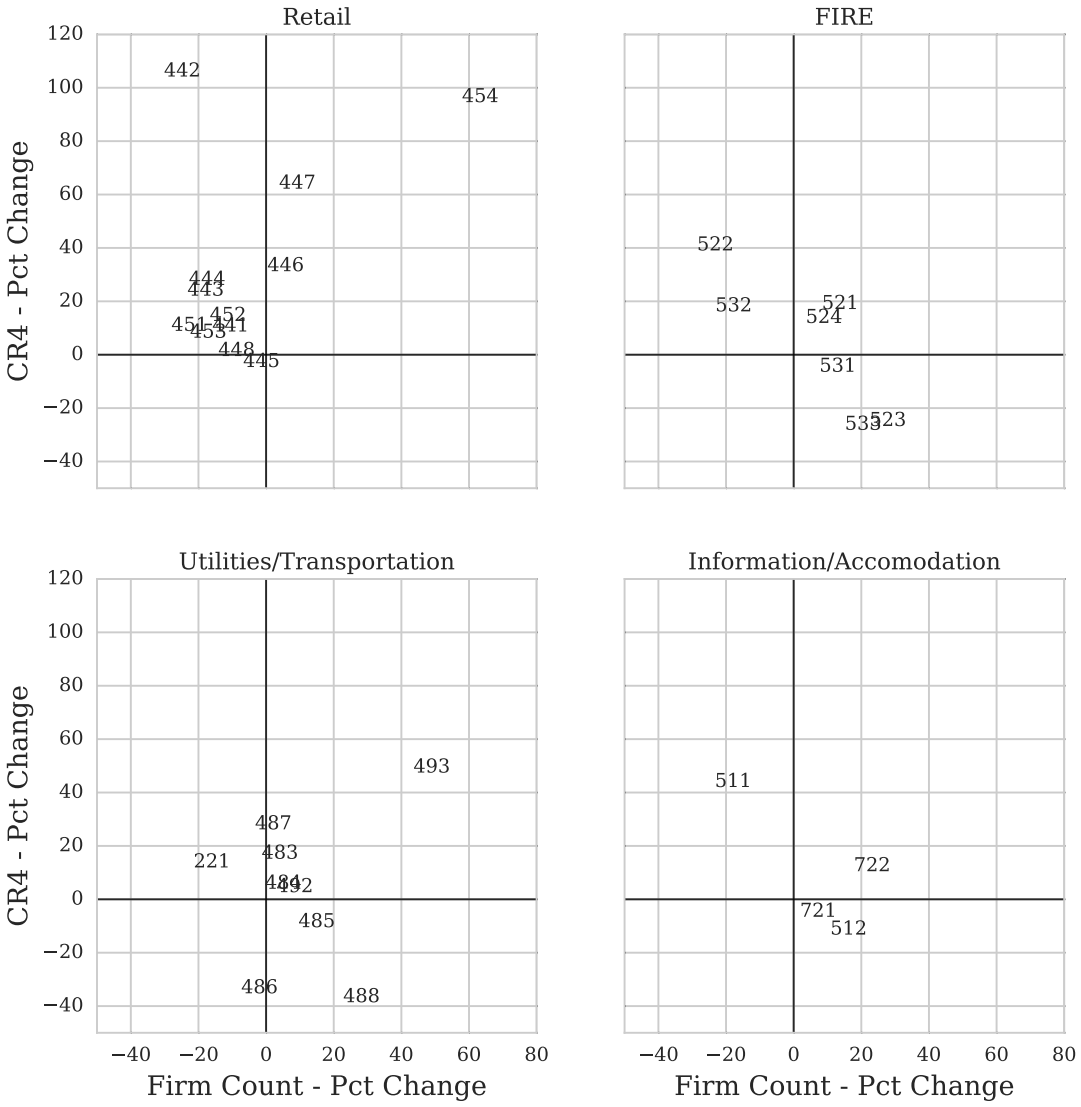
The consequences of increased market concentration within the US banking industry are, at least initially, beneficial to consumers of banking services. *Black and Strahan* (2002) and *Chava et al.* (2013) find small-business have increased access to capital as a result of removed interstate banking restrictions and increased competition. *Cetorelli and Strahan* (2006) support these findings, but also points out that a concentrated banking sector is detrimental for firms' access to capital. In this case, the immediate advantages of deregulation would diminish over time as the number of banks decreases. See *Berger et al.* (2004) for an extended survey on the consequences of increased market concentration among commercial banks.

2.3.2 Retail

US retail has evolved in dramatic ways over the past 15 years. In 2002, there were 11,265 firms selling books, periodicals, and music (NAICS code 4512). Of these, the four largest firms made up nearly 49% of industry revenue. In 2012, the were 4125 firms and the four largest firms account for over 66% of industry revenue (US census). Books stores, both large and small, point to e-commerce as a reason for their demise (*Herman* (2001), *Weisman* (2004), *Hooper and Rawls* (2014)). *Goldmanis et al.* (2010) finds that, due to how online classification of online retailers, it is difficult to confidently measure the impact of online sales on the success and survival of brick-and-mortar establishments. *Zhu* (2001) points to the internet as the primary reason for the collapse of the brick-and-mortar movie-rental business. Using phone book data, *Zentner* (2008) finds a causal relationship between the introduction of broadband Internet services and increased exit rates of music specialty stores.

The same trend exists, to varying degrees, for retailers selling furniture, electronics, building materials, and clothing. Figure 2.3 depicts the change in CR4 against the change in the number of firms between 2002 and 2012. For context, I also include panels for FIRE

Figure 2.3: Changes in CR4 vs Changes in Firm Counts by Industry Subgroups



Source: author's calculations from US Census data. The figure plots the percent-change in the total number of firms against the percent-change in the CR4 concentration ratio for 3-digit NAICS industry groupings between 2002 and 2012. The top-left panel plots NAICS categories for retail trade. The top-right panel plots NAICS categories for Finance, Insurance and Real-Estate. The bottom-left panel plots NAICS categories for utilities and transportation. The bottom-right panel plots for NAICS categories of information and accommodation.

(finance, insurance, real estate), utilities/transportation, and information/accommodation. Note that in the panel for retail, nearly all of the sub-industries are located in the upper-left quadrant with a positive change in CR4 and a negative change in the number of firms. This shows that between 2002 and 2012, a net positive number of retail firms exited and the largest firms increased their market share. The only exception is NAIC group that experienced a decrease in CR4 is 451, the group for food and beverage retailers. The only group to experience a large increase in the number of firms is NAIC group 454, which is the group for "nonstore retailers" which includes electronic stores. The decrease in the number of retail firms in most industries, then, is not because retail is becoming less popular. Retailers are simply adopting more "nonstore" business models which include electronic and mail order businesses that can compete compete across wider geographies.

The same cannot be said of other all other industries. In the panels for FIRE, information/accommodation, and utilities/transportation, NAIC groups experience a more balanced change in CR4. For FIRE and industry/accommodation, nearly half of the NAIC groups are *less* concentrated in 2012 than they were in 2002, as measured by CR4. In utilities/transportation, 3 of 10 NAIC groups experienced a decrease in market concentration while in retail only 1 in 12 NAICS groups experienced a decline in concentration. For changes in the number of firms, more than half of the non-retail NAIC groups demonstrate in increase in the number of firms between 2002 and 2012. While these non-retail industries experienced many changes between 2002 and 2012, they did not experience a dramatic market expansion as was observed in retail.

2.3.3 Trade liberalization

One simple way that markets can increase in size is by removing barriers to trade with neighboring markets. There are many types of policies which can accomplish this, including the removal of tariffs and quotas, entering into trade agreements or economic unions, or by developing new infrastructure to facilitate larger trade volumes. Here I detail a number of

well-studied policies and their affect on market size/concentration, measured in various ways over various economic studies.

The North American Free Trade Agreement (NAFTA), established in 1994, created a trade bloc between Canada, the United States, and Mexico. Over the next 10 years, nearly all remaining tariffs between the three countries were eliminated. In removing these barriers to trade, exporting firms simultaneously faced larger markets for selling their products and increased competition from imports. According to several studies, the years following NAFTA saw increased exit rates among smaller and more highly leveraged firms (*Head and Ries* (1999), *Gu et al.* (2003), *Baggs* (2005), *Colantone et al.* (2008), *Lileeva* (2008)). In addition to firm exit, *Breinlich* (2008) points to an increase in mergers and acquisition activity as a source of increased market concentration following Canadian-US trade liberalization.[3] The exit of firms in this setting does not necessarily suppress total output. As described in *Melitz and Ottaviano* (2008) and *Melitz* (2003), firm exit is simply the result of increased competition. In smaller markets with fewer competitors, firms can operate with some inefficiency and still remain profitable. When the market expands and new competition is introduced, resources will be reallocated to the firms that can produce most efficiently.

Increases in firm exit rates following the enactment of trade liberalization policy is certainly not limited to the US and Canada. Looking at US manufacturing plants and a cross-section of US trading partners, *Bernard et al.* (2003) estimates a 5% reduction in geographic barriers to trade would lead to a 15% increase in trade as well as the exit about 3% of firms. *Breinlich* (2008) finds results comparable to the NAFTA literature using data from Columbian manufacturing plants in the 1980's and 1990's. *Álvarez and Claro* (2009) also find similar results looking at Chilean manufacturing plants. In all of these studies, the removal of trade barriers resulted in larger markets which ultimately resulted in a more concentrated market as less efficient firms struggle in the face of increased competition.

[3]See *Long and Vousden* (1995) for another example of trade liberalization preceding an increase in mergers.

2.3.4 US Breweries and Mass Advertising

According to *Gokhale and Tremblay* (2012) and *Swaminathan* (1998), in 1935 there were about 766 major breweries in the US. Of these, the largest four firms accounted for about 11% of total sales. In 2012, there are only about 20 breweries and the largest four firms collect over 90% of total sales. Several studies, including *George* (2009), *Lee and Tremblay* (1992), and *Färe et al.* (2004) point to nationalized television campaigns and improved delivery and logistics that allowed firms to increase the potential size of their market. As the size of the market increased, the additional competition forced competitors to exit or merge with other firms. *Bhuyan and McCafferty* (2013) shows that the dramatic change in market structure affects the profitability of these firms.

An important response to this increased market concentration is the emergence of a thriving micro-brewing industry (*Tremblay et al.*, 2005). As described in *Cabras and Bamforth* (2016), local breweries are able to enter and successfully compete with national incumbents by offering a product that differs in some meaningful way. By inserting a degree of differentiation (distance) between themselves and their competitors, these firms are able to recover some of the advantages they experienced in a smaller, more isolated market.

2.3.5 Broad Trends in US Firm Dynamics

In addition to the specific examples mentioned above, there is a separate literature showing that since at least the 1970's the US has trended toward a distribution of larger and older firms. In other words, the US economy as a whole is becoming increasingly concentrated. *Decker et al.* (2013) documents this trend and shows that over several decades the trend is evident in all geographies and industries. *Reedy and Strom* (2012) and *Pugsley and Sahin* (2015) show the shift can be attributed to a declining startup rate. *Davis and Haltiwanger* (2014) document a decreased frequency of IPOs and decreased labor-market mobility. *Molloy et al.* (2016) claims that the trend is only partly explained by concurrent demographic trends and argues that a substantial portion of the trend is still unexplained. My claim is

that market expansion would facilitate this increase in concentration.

2.4 Model

Time is discrete, infinite, and indicated by t. The number of active firms is given by N_t. Each firm has a location denoted by $x_i \in [0,1]$, which allows for some notion of distance between firms in the spirit of *Hotelling* (1929). Distance can be interpreted geographically, but can also be thought of more broadly as differences in product offerings or customer populations.[4] For brevity we simply refer to all visible differences between firms as "distance" and assume firms all create the same product. The distance between firms i and j is given by the distance metric $d(x_i, x_j)$, which satisfies the mathematical properties of a distance metric.

Distance between firms is important in determining each firm's market-clearing price. Prices are set via an inverse demand function that parametrically accounts for the output of all other firms, weighted by distance.

$$p_{i,t} = \left(\frac{N_1}{N_t}\right)\left(\chi - \psi \sum_{j=1}^{N} w(x_i, x_j; h) q_{j,t}\right). \tag{2.1}$$

Here, ψ is the slope of the inverse demand curve. The weighting function $w(x_i, x_j; h) : [0,1]^2 \to [0,1]$ is non-negative, symmetric, and integrates to 1. Under these assumptions, $w(x_i, x_j; h)$ belongs to the generic family of kernel functions used in many statistical applications. The parameter h specifies the distance at which firms no longer are able to influence each other. As this parameter increases, firms find themselves subject to competition from increasingly distant firms. The term $\frac{N_1}{N_t}$ scales market share according to the number of active firms, such that surviving firms acquire a share of the market deserted by exiting firms and lose market share when a new firm enters. Due to this term, the size of the market is

[4] *Capozza and Van Order* (1982) shows that product differentiation within monopolistic competition models can be achieved using spatial distance. While a single dimension is sufficient for my purposes, others studies, such as *Phlips and Thisse* (1982), allow for multiple dimensions of product differentiation.

unchanged by the number of participants (See Appendix B.3 for additional details).

This pricing function implicitly assumes a distribution of households making decisions regarding which of several firms to transact with. As with a standard Cournot oligopoly model, it is not necessary to model the household explicitly as the sum of all preferences, constraints, and environmental conditions affecting the households are captured in the inverse demand function. The fact that my model generates a distribution of prices is justifiable in the presence of distance as a friction. In a frictionless environment, no household would optimally transact with a firm selling output at a higher price than another firm and the distribution of prices would collapse to a single market-clearing price. In my environment, however, geographic distance alters household decisions such that households can optimally transact with a firm even if its price is not the lowest.

2.4.1 Production and Profits

At the start of each period, each firm observes its private capital stock k_t as well as an idiosyncratic productivity shock z_t. Productivity shocks follow an AR(1) process.

$$z_{t+1} = \rho_z z_t + \epsilon_t \tag{2.2}$$

Output is given by the following production function:

$$q_t = z_t f(k_t) \tag{2.3}$$

where $f(\cdot)$ satisfies $f' > 0$, $f'' < 0$, and $f(0) = 0$. As all inputs are observed at the start of the period, output within the period is deterministic. Firms can only influence future earnings profits by choosing how much new capital to invest each period, denoted by i_t. Capital depreciates at rate δ and evolves according to a standard law of motion

$$k_{t+1} = k_t(1-\delta) + i_t, \tag{2.4}$$

$$\tag{2.5}$$

Firms face a per-unit cost c to manufacture a unit of output. Additionally, firms must pay a fixed cost c_m each period to stay open. The fixed cost can be thought of as overhead expenses, and is helpful for creating an environment in which firms will potentially choose to exit.[5] The firm's profit period t is consequently given by

$$\pi_t = p_t q_t - c q_t - i_t(1+r) - c_m \tag{2.6}$$

Here r is an exogenous interest rate that applies when revenue net of operating costs is negative and hence borrowing is necessary. I set r to a high value such that borrowing is possible but rarely optimal. The presence of a "soft" borrowing constraint is computationally convenient, but will also allow new firms to successfully enter without an existing capital stock.

2.4.2 Expectations over Variables Associated with Neighboring Firms

In a rational expectations setting, each firm would have to know both the productivity and capital holdings of all other firms in order to accurately predict future states of the world. This is clearly intractable (and unrealistic) for more than a small number of firms. Here, I focus on a boundedly rational setting in which a firm only uses weighted aggregate output as a state-variable. Let Q_i be the weighted sum of output from firms neighboring firm i.

$$Q_i = \sum_{j \neq i} w_j(x_i, x_j; h) q_j \tag{2.7}$$

[5] This is a standard feature in many labor-market models with endogenous firm exit. See *Hopenhayn and Rogerson* (1993).

How this variable will evolve over time is not at all clear. When the number of neighboring firms is small, or when a neighboring firm has a relatively high weight aggregate output Q_i will be highly correlated to the output of the closest neighboring firm. When a firm has several equally sized neighbors, it's possible that Q_i could stay close to the mean of its stationary distribution. It's also possible for the volatility of Q_i to change as firms enter or exit. Potentially changing volatility rules out simple forecasting rules as in *Krusell and Smith* (1999) as well as constant values for Q_i as in models like *Huggett* (1993) or *Hopenhayn and Rogerson* (1993). Consequently, I adopt a naïve expectation rule under which a firm observes $Q_{i,t}$ and expects that the value of $Q_{i,t+1}$ will be the same

$$E\left(Q_{i,t+1}\right) = Q_{i,t}. \tag{2.8}$$

As it is present in equation 2.1, the share of surviving firms $\left(\frac{N_t}{N_1}\right)$ is another state variable associated with neighboring firms. Since N_1 is fixed, this leaves N_t as the only variable firms need to observe each period. As I do with output, I assume firms only observe the share of surviving firms instead of tracking the activity of individual firms (even firms in close proximity). Firms assume the future value of N_t is the same as observed within the period.

$$E\left(N_{t+1}\right) = \left(N_t\right). \tag{2.9}$$

These simplifying assumptions make this model solvable, but also remove opportunities for strategic policies as in *Guillén and Pinto* (2007). Under naïve expectations, a firm ignores the interaction between its decisions and the the decisions of other firms which may not be realistic when the number of firms is small. While this assumption simplifies my equilibrium concept considerably and allows the model to be solvable even for large values of N, it comes at the expense of a rich set of collusive or otherwise strategic policies. There are methods for solving dynamic models with heterogeneous firms, such as those presented in *Ericson and Pakes* (1995) *Pakes and McGuire* (2001) for finding Markov Perfect equilibria.

These solution methods, however, do not scale with the number of firms. To overcome this, *Weintraub et al.* (2008) proposes a new equilibrium concept which allows firms to remain oblivious to the states of other firms and focus instead on the aggregate states of the economy – even in a strategic setting. Unlike *Weintraub et al.* (2008), however, firms in my model do not necessarily perceive the same aggregate state of the economy ($Q_{i,t}$, while an aggregate variable, differs across firms due to their different locations and proximities to other firms). Like *Krusell and Smith* (1998) and *Krusell and Smith* (1999) firms make their decisions based on aggregate observables which allows their model to include a large number of firms without increasing the dimensionality of the state space. The key characteristic that distinguishes these boundedly rational approaches from more traditional full-information approaches is that boundedly rational solution methods base expectations on aggregates or patterns which summarize available data rather than keeping track of every data point individually. My solution method follows this boundedly rational tradition by assuming firms focus on simple aggregate statistics $Q_{i,t}$ and N_t rather than attempting to monitor an ever-increasing state space as the number of firms increases (an assumption that is both intractable and unrealistic for a large number of firms).

It is important to note that forming expectations in this way, while computationally important, can result in wildly cyclical short-term behavior if those expectations are ever wildly incorrect. In my simulations in Section 2.5.3 I push the model to these extreme scenarios by dramatically changing the parameter h over a short number of periods. These cycles still occur around clearly discernable long-run trends, however, and these long-run trends are my primary interest as changes in market concentration typically take place over several decades. To ensure the model had more realistic short-term dynamics, one would only need to form expectations in a way that considers a longer-term average value for $Q_{i,t}$ and N_t instead of simply taking the previous values and ignoring the previous history.

2.4.3 Entry

Firm entry is modeled in a fashion comparable to *Hopenhayn and Rogerson* (1993). At the start of each period, each firm in a pool of potential entrants draws a location x_i along with an idiosyncratic productivity shock z_t. As neighboring firms produce, potential entrants can observe the output of neighboring firms and all other variables observed by active firms, Q_i and N_t. At the end of the period, potential entrants decide whether to pay a one-time entrance fee c_e and become an active firm. The firm has no capital, but can borrow to acquire i_t which will enable production the following period. The new entrant must also pay the operating fee c_m, required by all active firms. New entrants will not produce within the period, as their period t capital holdings are zero. Consequently, entering firms knowingly incur a loss within the period but expect to earn positive profits in the future.

Altogether, the within-period profits for new entrants is given by

$$\pi_e(i_t) = -i_t(1+r) - c_m - c_e. \tag{2.10}$$

Rather than track potential entrants over time, I assume potential entrants which choose to not to enter the economy exit the pool and are replaced by a new population of potential entrants. This is done without any loss of generality as the distribution of z_t shocks is stationary. The decision to enter is denoted by $\tilde{e}_t \in \{0, 1\}$, where $\tilde{e}_t = 1$ represents a decision to enter.

2.4.4 Exit

At the end of each period, after earnings are realized, a firm makes a binary choice concerning whether it will continue to operate. This decision is denoted by $e_t \in \{0, 1\}$. Due to the persistent nature of productivity and capital, as well as the fixed cost for remaining active, there are some states in which a firm will anticipate an extended duration of negative profits. If these negative profits are severe enough, a firm will find it optimal to exit. Existing

capital is discarded at no cost.

An important consequence of firm exit is that nearby surviving firms will realize higher prices and profits for the same level of output. The exiting firm no longer enters into the inverse demand function of the surviving firms. As firms continue to exit, the surviving firms will come closer to targeting monopolistic levels of output. As this happens, the departure of firms should decelerate and the economy should reach a stable distribution of firms.

New entrants will also make an exit decision, but will never actually choose to exit as they have no new information at the end of the period as they did at the start. Any firm with a high enough z_t value to enter cannot simultaneously expect an extended future of losses low enough to motivate an exit.

2.4.5 Dynamic Optimization

The firm's problem is to maximize expected profits over its lifetime, which is potentially finite if the firm chooses to exit. Firm's face no cost after exiting, so the firms value function value upon exit, denoted by J_{exit}, is simply 0.

$$J_{exit} = 0 \tag{2.11}$$

The value function for active firms, given by J, includes profits for the current period in addition to a continuation value. If a firm remains active by choosing $e_t = 0$ the continuation value will be given by J. If the firm chooses $e_t = 1$, the continuation value will be equal to J_{exit}. Altogether, the firm's problem is given by

$$J(z_t, k_t, Q_t, N_t) = \max_{i_t, e_t} \pi_t(z_t, k_t, Q_t, N_t) + \beta \left((1 - e_t) E_t \left[J(z_{t+1}, k_{t+1}, Q_{t+1}, N_{t+1})\right] + e_t J_{exit}\right)$$

subject to the following set of constraints

$$E\left(Q_{i,t+1}\right) = Q_{i,t} \tag{2.12}$$

$$E\left(N_{t+1}\right) = N_t \tag{2.13}$$

$$k_{t+1} = k_t(1-\delta) + i_t \tag{2.14}$$

$$z_{t+1} = \rho_z z_t + \epsilon_t \tag{2.15}$$

The firm's problem for potential entrants differs slightly. Let J_e denote the value function for potential entrants and $\tilde{e}_t \in \{0,1\}$ the decision to enter or not. When a potential firm enters, $\tilde{e}_t = 1$ and the new firm receives $\pi_e(i_t)$ within the period and J as a continuation value. When potential entrants who do not enter, $\tilde{e}_t = 0$ and the firm leaves the pool of potential entrants with a continuation value of zero, J_{exit}

$$J_e(z_t, k_t, Q_t) \quad = \quad \max_{i_t, \tilde{e}_t} \tilde{e}_t(\pi_e(i_t) + \beta E_t\left[J(z_{t+1}, k_{t+1}, Q_{t+1})\right]) + (1-\tilde{e}_t)J_{exit} \tag{2.16}$$

subject to $k_t = 0$ as well as (2.8), (2.9), (2.4), and (2.2).

2.4.6 Equilibrium Concept

As a result of naïve expectations, a single firm's actions have no expected impact on its neighbors. Consequently, I can ignore strategic interactions between firms and make use of a simple recursive equilibrium concept.

Definition II.1. For a set of variables k_t, Q_t, and z_t, a recursive equilibrium is defined as a set of policy functions $e_t(k_t, Q_t, z_t)$, $\tilde{e}_t(k_t, Q_t, z_t)$, $i_t(k_t, Q_t, z_t)$ and firm-specific price functions $p_t(q_t, Q_t)$ such that both active firms and potential entrants maximize their lifetime expected earnings $J(z_t, k_t, Q_t)$ subject to equations (2.8), (2.9), (2.4), and (2.2).

Verifying the existence of an equilibrium is straightforward since the inverse demand curves are taken as given for each firm. Since the production function is concave in capital

and marginal costs increase linearly with output, there must exist an optimal level of capital that would maximize long term profits. Since firms form naïve expectations for total output in their neighborhood, firms cannot anticipate the consequences their actions may have on the firms around them. Consequently, there are no strategic complexities to consider. If firms developed forward looking expectations for aggregate output in their neighborhood, however, the model would become a repeated game in which firms may attempt to overproduce in the short-run to drive out competition and achieve long-term gains.

2.5 Quantitative Results

Here I provide a brief overview of my solution method. A detailed description of how to solve the model is included in Appendix B.1. Additionally, all code and programs used in solving the model are included with documentation at `https://github.com/btengels.com`.

2.5.1 Model Calibration

Since firm location is projected onto a single dimension, I use a simple distance function $d(x_i, x_j) = |x_i - x_j|$. This distance function enters the model only through the weighting function w (equation (2.1)). Distant firms will have a proportionally smaller impact on a firm's price in comparison to competing firms in close proximity. The functional form I choose for w is a triangular kernel function. With this functional form, competitor weights linearly diminish with distance until a certain maximal distance is reached. The results were found not to change in any important ways under alternative kernels such as a parabolic weighting function. The equations for this weighting functions are included in Table 2.1, and Appendix B.3 includes details how the resulting weights vary with the bandwidth parameter h. Weights are weakly positive, such that if a competitor for a given firm is too distant the resulting weight is zero.

In order to maintain focus on the novel model ingredients, the remaining elements and assumptions of the model are set to values typical to macroeconomic studies. The produc-

Table 2.1: Model Calibration

	Value
FUNCTIONAL FORMS:	
Production function	$f(k_t) = k_t^{\alpha}$
Distance function	$d(x_i, x_j) = \lvert x_i - x_j \rvert$
Competitor Weighting Function	$w(x_i, x_j; h) = \max\left\{0, \left[1 - \frac{d(x_i,x_j)}{h}\right]\right\}$
MODEL PARAMETERS:	
Discount factor β	0.98
Depreciation rate δ	0.02
Production function curvature α	0.5
Productivity autocorrelation ρ_z	0.98
Productivity volatility σ_z	0.05
Slope of inverse demand function ψ	0.5
Marginal cost c	0.1
Operating cost c_m	0.4
Entry cost c_e	0.3
Initial number of firms N	100

tivity of capital is given by $f(k_t) = k_t^{\alpha}$ with $\alpha \in (0, 1)$. I assume each period corresponds to a quarter and consequently set the discount factor β to .98 and the capital depreciation rate δ to .02. As my model has no exogenous aggregate uncertainty, I choose parameters for the evolution of z_t such that the idiosyncratic volatility of firm output is similar to *Hopenhayn and Rogerson* (1993).[6]

2.5.2 Model Solution

To find the optimal policy functions, I set up a discrete state space for productivity z, capital k, and the weighted sum of competing output Q. I compute the transition probabilities between different productivity states using the method outlined in *Rouwenhorst* (1995). I similarly approximate the choice variable i using a discrete grid (choice variable e_t is already discrete). See Appendix B.1 for a more detailed description of my solution method,

[6]Aggregate uncertainty still exists in the model due to firm entry and exit. This, however, is entirely endogenous.

as well as parameters governing idiosyncratic productivity for firms.

At each possible state, I compute the period t profits for each possible (z, k, Q, i) combination. On this feasible set, I use value function iteration to solve for the values of J satisfying the firm's problem.

2.5.3 Model Simulations

I solve and simulate the model under 3 main scenarios, detailed in Table 2.2. Simulations feature 100 initial firms with locations randomly drawn over the $[0, 1]$ interval. Simulations begin after a sufficiently long burn-in period and continue for 300 periods after burning in. Each parameterization is simulated 5000 times so as to generate a comprehensive range of outcomes.

Each of the three scenarios explores a different path for h over time. In every simulation, however, h begins at a very small value of 1e$-$8, which is small enough that firms act as local monopolists. Then h transitions to larger values of .01, .03, and .05, respectively, in the first, second, and third scenarios. In the first case, h quickly jumps to these higher values in imitation of a legal or policy changes occurring over 25 quarters. In the second case, h increases linearly over 60 periods. In the final case, h increases gradually over 150 periods, in imitation of demographic or technological changes which increase market size gradually over several decades.

Table 2.2: h values in simulations

	Description	Final value
Scenario 1	Linear increase (25 periods)	.03
		.06
		.1
Scenario 2	Linear increase (60 periods)	.03
		.06
		.1
Scenario 3	Linear increase (150 periods)	.03
		.06
		.1

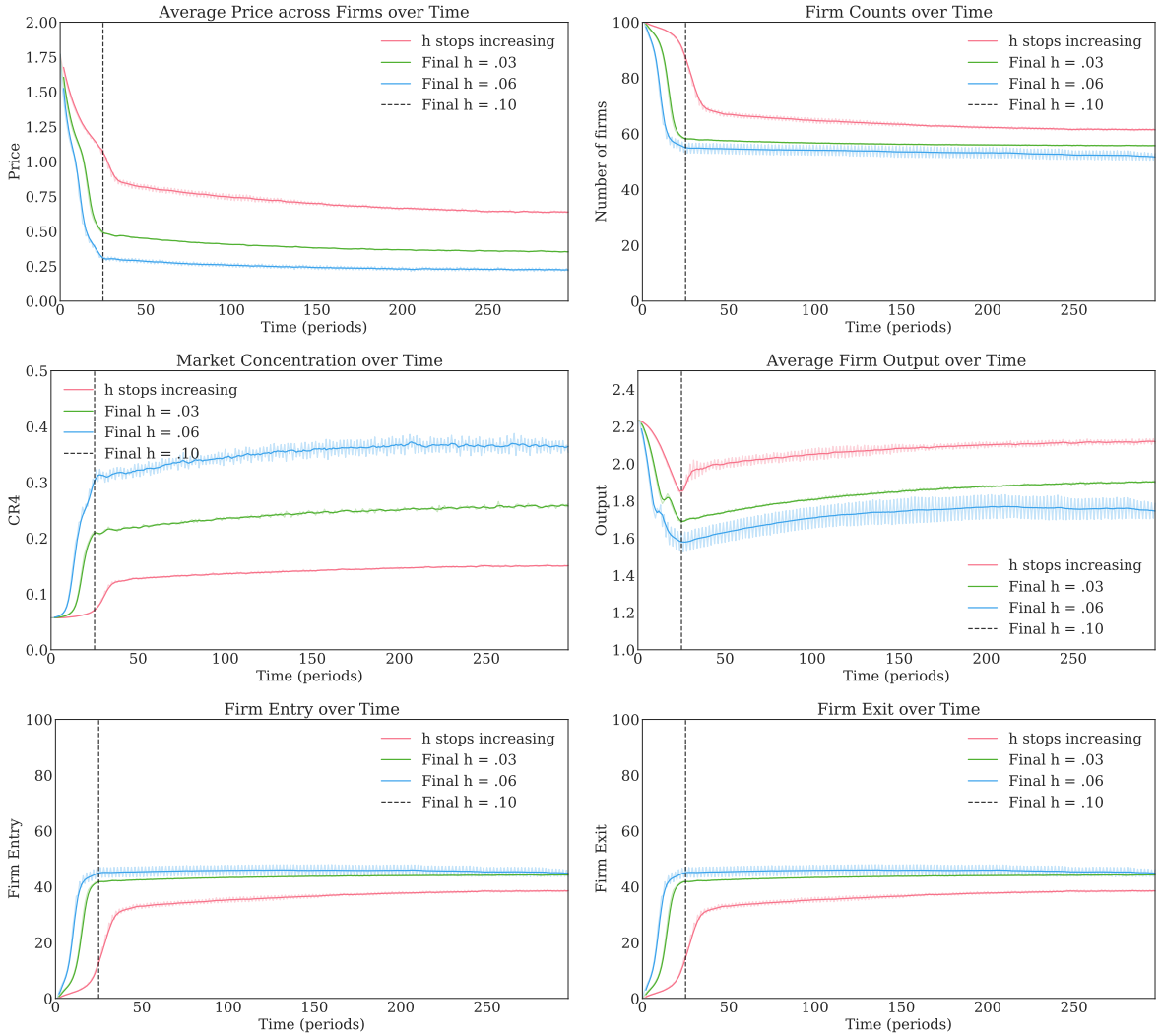
The results for the first scenario are shown in Figure 2.4. Short term movements are denoted in light colors while long-term trends, my primary focus, are denoted in darker colors. Initially, all firms act as monopolists and share a common price. Then prices drop rapidly from $t = 1$ to $t = 25$ as h increases and firms face increased competition from neighboring firms. Simulations with larger values of h demonstrate more severe decreases in price. Once h stops increasing at $t = 25$ prices change much more slowly, decreasing slightly over the final 250 periods.

As prices decrease, so does the number of active firms. In simulations where h increases to .03, the number of active firms stabilizes near 65. Larger increases in h result in fewer active firms, with an average of 60 and 55 active firms resulting from h increasing to .06 and .1, respectively. Once h increases, the number of firms begin to oscillate from one period to the next. The severity of these oscillations is greater for larger increases in h, and is a natural consequence of the backward-looking expectations in (2.9). Since all active firms and potential entrants observe the same number value for N_t and expect the same value in the following period, they are all surprised the following period to find more (less) active firms the following period than expected. As firms expect this new value of N_t for the following period, some firms will exit (enter) and the economy will return to the state it was two periods prior. When h increase by small amounts between periods, these naïve expectations are correct enough to avoid any kind of hog-cycle. Alternative methods of forming expectations (e.g. moving average or some linear function fit to historic values) would yield more stable results. Even with these oscillations, however, the long-term relationship between market size and market concentration is still apparent.

Firm entry and exit rates are substantial, especially for simulations where the final value of h is large. In the long-run when there are about 60 active firms, nearly 40 firms are exiting/entering the economy each period. Only firms with very fortunate positions x_i and productivities $z_{i,t}$ survive for more than a few periods.

Average output per firm steadily decreases while the economy transitions to higher values

Figure 2.4: Scenario 1: Increase in h over 25 quarters



Lightly colored lines are the average period t value across all simulations. The darker colored lines are a 3 period centered moving average and denote the long-term trends which are my primary focus. The vertical dashed line denotes when h finishes expanding.

of h. Once h stabilizes, however, and both prices and the number of active firms reach a more consistent range of values, output per firm begins to steadily increase as firms scale up to serve larger markets. Overall, the changes in output per firm are minor compared to the large changes observed in price and in the number of active firms.

Finally, market concentration as measured by CR4 rapidly increases as h increases. In a monopolistic setting, the four largest firms account for about .04 of total output (there are

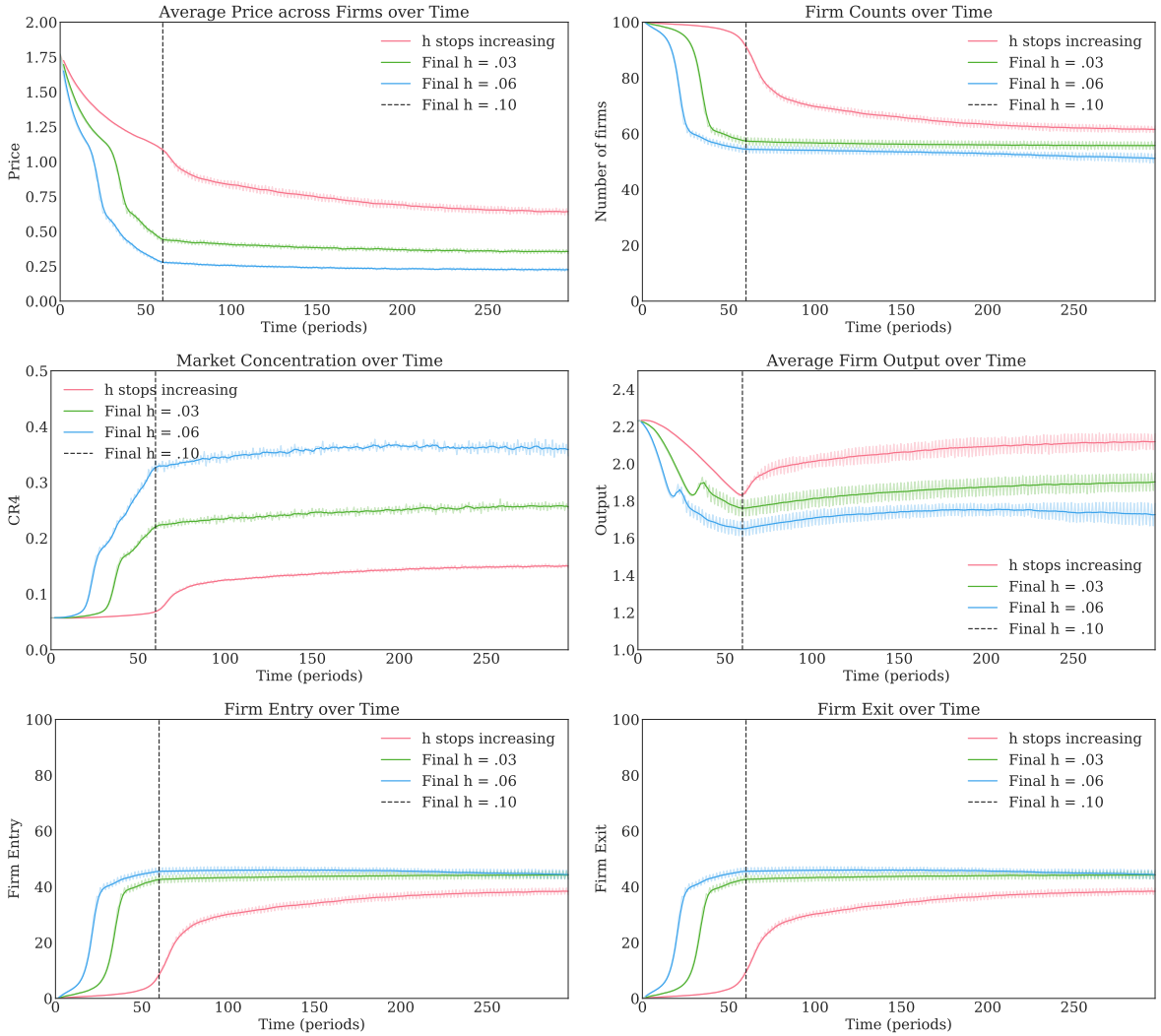
100 firms initially). In these simulations, CR4 increases to values of .13, .25, and .35, as h increases to .03, .06, and .10, respectively. Some of this is the natural consequence of having fewer firms. If all active firms were approximately the same size, however, the resulting CR4 values for these simulations would be between .06 and .07 given the long-run number of active firms. A CR4 value of .35, then, suggests that some firms have become disproportionately large following an increase in market concentration while other firms continue to operate at a much smaller scale.

In the second scenario, shown in Figure 2.5, h begins at monopolistic values of $h = 1e{-}8$ and linearly increases to values of .03, .06, and .010 over a longer time frame of 60 quarters. While the simulations are similar in many ways to those in the first scenario, there are also several differences. One notable difference is the smaller oscillations in the number of active firms. This is because the slower changes in h allow firms to more correctly form expectations over the number of active firms. As in the first scenario, increasing h decreases prices as markets become more competitive. This leads to an increase in firm exits and market concentration. Even though h increases less rapidly than in the previous case, the long-run values for endogenous values are mostly the same as in scenario one. Importantly, the average CR4 within my simulations for this scenario is several times higher than it would be if all firms were approximately the same size at the end of the simulation.

The results from the third and final scenario are shown in Figure 2.6. This scenario has the same initial values and ending values for h as the other scenarios, but in this scenario h increases slowly to its final value over the course of 150 periods. This is in imitation of markets increasing from slow changes in demographics or taste.

As before, the long-run implications of an increase in h are the same as the previous scenarios with a few key differences. With smaller changes in h from one quarter to the next, expectations over N_t are more accurate and the model avoids large oscillations over the number of active firms. Another trend that becomes more apparent when h increases slowly is that even though h increases linearly, the response in prices is not similarly linear

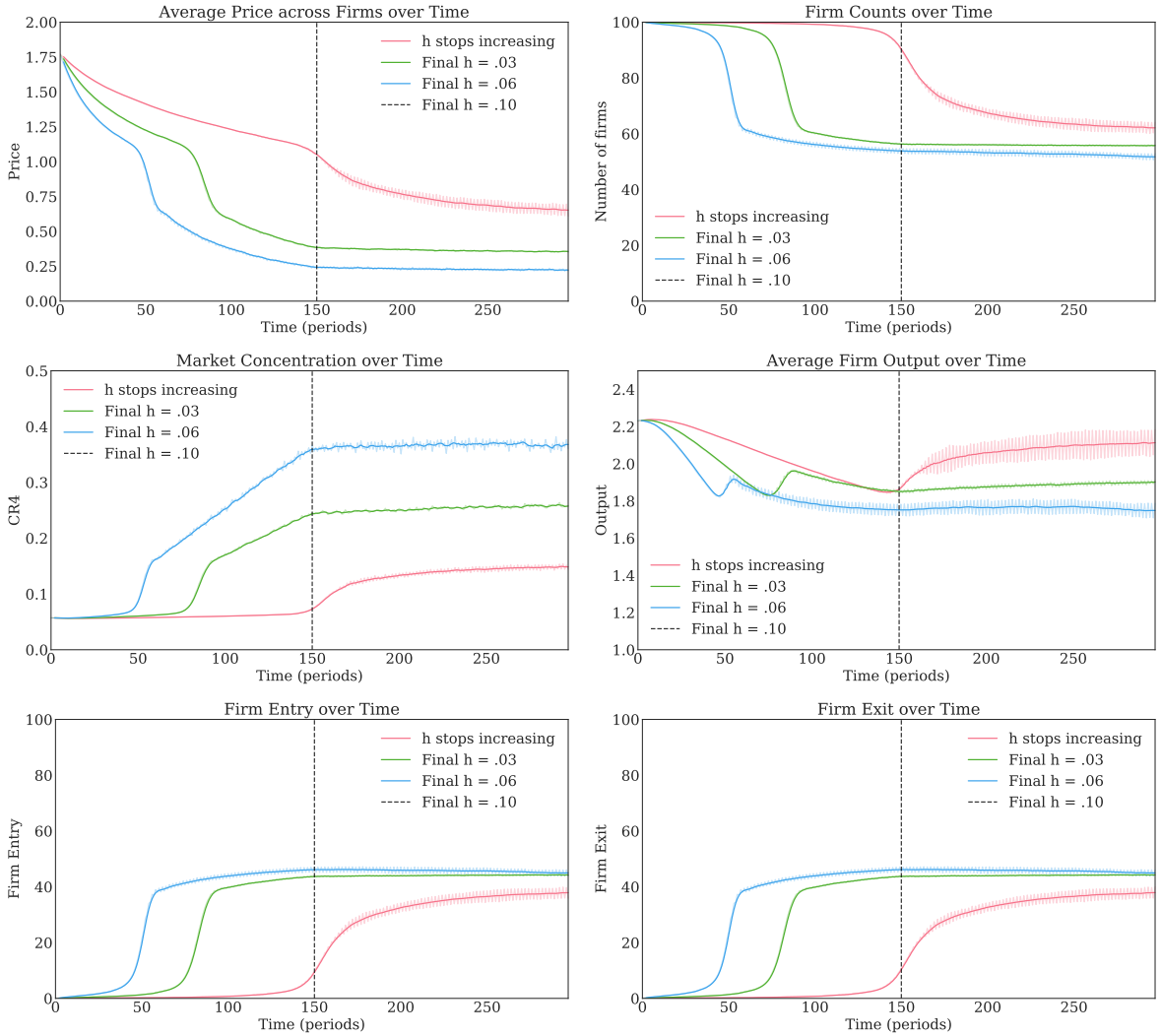
Figure 2.5: Scenario 2: Increase in h over 60 Periods



Lightly colored lines are the average period t value across all simulations. The darker colored lines are a 3 period centered moving average and denote the long-term trends which are my primary focus. The vertical dashed line denotes when h finishes expanding.

if h is sufficiently large. When h is increasing to it's largest value of .10, prices decrease gradually for the first 50 periods, then rapidly decline over the next 5 periods, and then resumes a more gradual pace of decline until h stabilizes and prices hold steady at their long-run values. Finally, a key difference between this scenario and the previous scenarios is that the long-run average output per firm varies less for different values of h than in previous scenarios where h increased more rapidly.

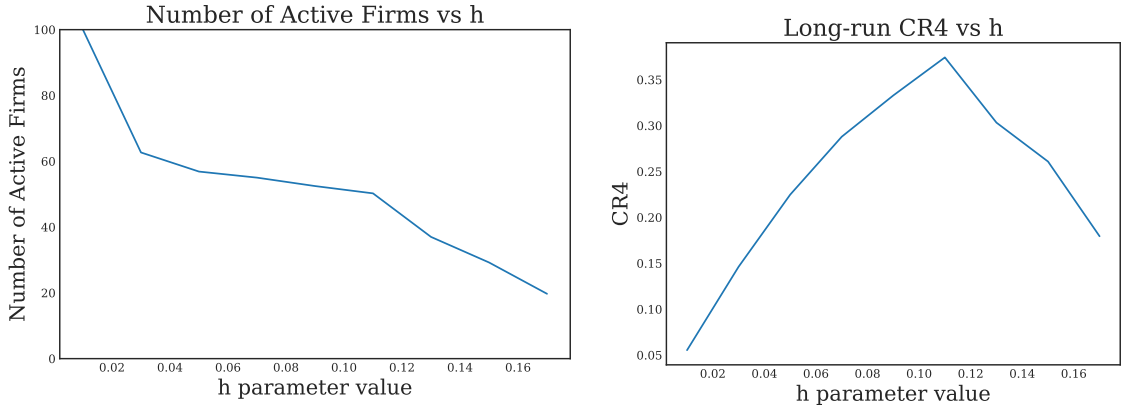
Figure 2.6: Scenario 3: Increase in h over 150 Periods



Lightly colored lines are the average period t value across all simulations. The darker colored lines are a 3 period centered moving average and denote the long-term trends which are my primary focus. The vertical dashed line denotes when h finishes expanding.

As a final exercise, I examine the long-run values of CR4 and the number of active firms over a greater range of terminal h values. The results are shown in Figure 2.7 I find that while the number of active firms decreases monotonically as h increases, market concentration as measured by CR4 does not similarly increase. Instead, CR4 increases with h up until $h = .11$ at which point it begins to decrease with h. Even though there are fewer firms, the distribution of output across firms is more uniform for $h > .11$. This is because

Figure 2.7: Key Long-run Values vs h



Long-run values are computed as an unweighted mean of the final 10 periods across 70 simulations.

larger values of h tend to decrease prices. Initially, these decreased prices force inefficient firms to exit. The decreased prices, however, also mean that the largest firms will produce less which pushes CR4 downward. As currently calibrated, the affect of decreased output dominates for values of $h > .11$. This trend would eventually reverse if h continued to increase as CR4 necessarily approaches 100 as the number of active firms tends to 4.

2.6 Model Interpretation

While the interpretation of h is clear in the context of this model, a realistic counterpart is not immediately obvious. Market boundaries are not readily observable in most industries and consequently difficult to measure or track over time. In light of these apparent difficulties, I recommend a few directions to consider in thinking of an empirical counterpart of h.

One way to measure the strength of competition between two firms is to examine their marketing efforts. If firms selling similar products are launching marketing campaigns in the same locations and are targeting the same populations, then they would be considered stronger competitors than firms with marketing campaigns in different locations. A simple metric which accounts for the relative sizes of marketing expenditures is the Jaccard distance

between the set of customers exposed to each marketing campaign.[7] Similar distances could be computed between sets of locations containing some physical presence of a firm or sets of customers who have made visits to or purchases from a firm. The average distance between firms over time would map to a value for h.

How this model applies to an industry will, of course, vary across industries. In Section 2.2 I described several industries where market expansion facilitated increased market concentration. In the context of the model in Section 2.4, each industry can be described in part by the parameter h which determines the extent to which firms are in competition with neighboring firms. The rate at which h changes differs across industries, as demonstrated by the long and slow evolution of the brewing industry as compared to the relatively rapid evolution of the US retail and banking sectors. Some industries such as utilities (see Figure 2.3), have no apparent change in market concentration in recent years. Goods and services in this industry are, by nature, expensive to transport large distances. More specifically, water and electricity are often sourced locally and have no "wireless" way of reaching customers as in retail and entertainment industries. While it is impossible to anticipate how future technologies might change these industries, to date technology has not expanded utility markets in dramatic fashion as seen in other industries. In the context of my model, new technologies cause h to increase within an industry only if the technology enables goods and services to reach customers more easily.

Finally, an econometrician must be careful to not simply call any increase in market concentration the result of market expansion. A merger between two firms, for instance, may have the effect of expanding markets and increasing market concentration simultaneously. In my model, the causality is evident. In any empirical exercise, however, establishing the direction of causality would be more difficult. One advantage of any specific industry-level analysis is that available data on entry and exit rates can help discipline parameters such as h, even if h is though to be changing over time.

[7] Jaccard distance is given by $1 - S(A, B)$ where $S(A, B) = \frac{A \cap B}{A \cup B}$.

2.7 Conclusion

It is difficult to understate the importance of understanding the causes and consequences of market expansion. The steady transition from a large collection of small markets to a small collection of large markets has important implications for financial industry, income inequality, housing prices, and entrepreneurship, among other things. Even so, the literature examining general patterns of market expansion is small in comparison to various industry-specific literatures. This is likely due to the difficulty of measuring market size, especially when changes occur gradually as the result of technological improvements or demographic changes. In the absence of readily-available data, I motivate my theory by identifying a common pattern which has played out across many different industries: as markets expand, increased competition results in a smaller number of firms and a more concentrated industry. This paper also takes a first step in developing a tractable, theoretical framework which can be used to frame future discussions on market expansion and its consequences.

The model adds the notion of distance within a Cournot-like pricing function such that nearby firms compete more aggressively than distant firms. Firms observe persistent productivity shocks as well as the aggregate competitive landscape as it affects their market clearing price. Firms make decisions each period regarding capital investments and whether or not to exit the economy entirely. In one model version, a firm's inverse demand function is invariant to the exit of other firms. This acts as a point of comparison to another version of the model in which inverse-demand functions shift out when firms exit to ensure aggregate demand remains unchanged. The model parameterizes the notion of market size, allowing me to examine the impact of market expansion on firm investment and exit decisions in model simulations.

In my theoretical setting, I find that market expansion increases competition and creates downward pressure on prices. Firm exit, on the other hand, reduces competition and consequently creates upward pressure on prices. In the version of the model when demand curves shift outward among surviving firms, the accumulation of demand among surviving

firms creates additional upward pressure. The relative strength of these competing forces determines whether prices ultimately increase or decrease. I find that the magnitude of market expansion, and not the pace of market expansion has an effect on long-run prices. The effect is parabolic in nature, such that mid-ranged levels of market expansion increase long-run prices while small and large amounts of market expansion result in lower long-run prices.

As this model is novel in its treatment of market expansion, there are many opportunities for future research. Here I detail three possible paths for future work. The first would be to consider the affect of market expansion on firm entry. In a model with market expansion, increased competition would likely reduce firm entry rates and create a net decline in the number of active firms. This would provide some theoretical underpinnings to the largely empirical literature on declining entrepreneurship rates (see *Decker et al.* (2013), *Decker et al.* (2014), and *Davis and Haltiwanger* (2014) for examples.) A second opportunity for future research would be to use observed exit rates within specific industries and use a model such as the one in this paper to back out the most likely path for h in order to match the observed exit rates. One could then compare the quantified histories of market expansion across industries and determine where the economy would be under a different history. Finally, a third opportunity for future research would be to introduce a degree of asymmetry in how firms experience market expansion. The introduction of an online retail platform, for example, will expand h for all firms in the industry, but will also transfer demand for output from the late adopters of the platform to the early adopters.

CHAPTER III

Group Punishments without Commitment

3.1 Introduction

Teams exist in many economic settings, ranging from teams of individuals working together in clubs or firms, to teams of companies in the form of cartels and lobby groups, to teams of nations in the form of political alliances and economic unions. In each of these settings, teams aim to improve outcomes by coordinating efforts across members and are often successful in doing so. Organizing as a team, however, may also introduce moral hazard problems, especially when team outcomes are shared and individual effort is not perfectly observed.

In static environments of team production, *Holmström* (1982) shows the only way to alleviate moral hazard problems is to rely on an outsider who can punish the entire team following a deviation from any team member. Punishments take the form of throwing away some share of the team's output. *Holmström* (1982) argues that the intervention of an outsider is also necessary to implement such punishments in a repeated environment, as the team might not want to enforce these punishments once team production outcomes are realized: *"There is a problem [...] in enforcing such group penalties if they are are self-imposed by the worker team. [...] Ex post it is not in the interest of any of the team members to waste some of the outcome. But if it is expected that penalties will not be enforced, we are back in the situation with budget-balancing, and the free-rider problem reappears."*

In this paper, we ask if and under what conditions outsiders are truly needed to enforce group punishments in a repeated context. In other words, we ask whether the ability of individual team members to punish other team members in the future enables the team to enforce group punishments which occur *after* aggregate outcomes are realized but *before* the realization of individual payoffs in the current period. We call such within-the-period punishments *static group punishments*. We show that, depending on the nature of the payoffs that agents obtain from team production, the team can indeed enforce static group punishments. In such cases, the threat of static group punishments is welfare enhancing relative to an environment in which the team's action set does not allow for static group punishments.

We start our analysis from a generalized model of repeated team production, featuring a team of agents and a benevolent Principal—a construct to represent team-wide preferences. In our model, agents individually choose a level of effort to contribute to the realization of a common outcome. After observing this common outcome, the Principal chooses a group punishment (possibly zero) which negatively affects the common outcome. The Principal, like the agents, cannot commit to a long-term strategy for group punishments. Since the Principal's action occurs after the common outcome is observed, the benevolent Principal values period utility of all agents plus the sum of future discounted stage-game payoffs of all the agents.

Our main contribution is to show that a broad class of repeated team production environments admits a simple, recursive characterization for the set of perfect-public equilibria. Specifically, we show how to characterize the entire equilibrium set of our generalized team production model using simple "carrot-and-stick" strategies for the worst perfect-public equilibrium (as in *Abreu* (1986)). We show that group punishments reduce the gains from deviations in the "carrot" phase, but increase the gains from deviations in the "stick" phase. Therefore, deviations from the "stick" never call for the implementation of group punish-

ments, further simplifying the recursive characterization of the equilibrium set.[1]

Our main findings are that static group punishments can be enforced by the threat of future actions by *team members*; and that the threat of static group punishments strictly improves the best attainable equilibrium welfare relative to an economy where the Principal's actions are restricted to never implement group punishments. Moreover, we show that a necessary condition for static group punishments to improve welfare is the presence of complementarities between aggregate outcomes and private actions in team members' stage game payoffs. We show that the *total static deviation payoff* (the total payoff that a deviant team member obtains within the deviation period) can be expressed as the deviant's static private gain minus a cost to incentivize the Principal to implement group punishments. Absent complementarities between aggregate outcomes and private actions, group punishments have no impact on this total static deviation payoff, and are therefore ineffective in deterring individual deviations—an outsider à la *Holmström* (1982) is required to improve welfare. Conversely, when team members' private actions interact with aggregate outcomes group punishments do reduce the total static deviation payoff by indirectly reducing team members' private incentives to deviate. In these cases, group punishments are useful to deter individual deviations, and an outsider may not be needed to improve welfare.

Our findings in the generalized model indicate that in presence of complementarities between aggregate outcomes and private actions, the Principal who lacks commitment (i.e. the team) might be capable of replicating incentive schemes which do not satisfy budget balancing without the aid of outsiders. In the second part of the paper, we apply our generalized team production model to the repeated oligopoly model of *Abreu* (1986), and ask which features of producers' payoffs make self-imposed group punishments most effective in improving team welfare—and therefore limit the need for an outsider. In the oligopoly model, team members are producers individually choosing how much output to produce, and the team outcome is the common price faced by all producers (a decreasing function of

[1] In the literature review, we argue that imperfect observability plays a key role in our recursive characterization, making continuation payoffs independent of the identity of the deviator (*Mailath et al.* (2017)).

aggregate team output). On the other hand, the group punishment imposed by the Principal is a tax rate (possibly zero) which has the effect of reducing the price of producers' output. As in the generalized model, the Principal cannot commit to a long-term strategy for taxes.

Within the context of the oligopoly model, we first show that group punishments imposed by the Principal are particularly effective in increasing team welfare for intermediate levels of the producers' discount factor. Intuitively, when producers are very impatient the threat of future punishments is weak and only small group punishments can be sustained following static deviations. For intermediate levels of the discount factor, the team can sustain large enough static group punishments such that the threat of these punishments allows the team to achieve the socially-optimal level of production. When producers are very patient, the threat of future punishments is strong enough that the team can sustain the socially-optimal level of production even without resorting to group punishments. Second, we show that for a given level of the discount factor group punishments are more effective when producers' output is highly substitutable. In these cases, deviations by individual producers have a small impact on the common price, increasing producers' static incentives to deviate, and increasing the ability of group punishments to improve team welfare relative to an economy where group punishments are not part of the team's action set.

Related Literature Our paper is related to a large literature concerning moral hazard in static team production settings. *Alchian and Demsetz* (1972) describe the opportunity for team members to shirk and still receive compensation and the need for a principal to prevent shirking. *Holmström* (1982) suggests a particular kind of contract in which a principal withholds payment whenever output is below its socially optimal level. Other studies solve the moral hazard problem by injecting a degree of competition among team members via tournaments, rankings, or other relative performance measures (see *Hart and Holmström* (1986) for a survey).

One of the main challenges in taking these static team production games to the infinitely-

repeated domain is to characterize the set of perfect-public equilibrium payoffs. *Mailath et al.* (2017) show that in a wide range of extensive-form games (including team production games) the equilibrium set cannot be characterized using simple penal codes, because both within-period punishments and continuation payoffs need to fit the identity of the deviator after a deviation has occurred. In our paper, we assume that group punishments can only affect team outcomes (due to imperfect observability), and show how under this assumption the equilibrium set can be characterized using simple penal codes. In other words, we show that simple penal codes can be used to characterize the entire set of perfect-public equilibrium payoffs in a broad set of repeated extensive-form games featuring imperfect observability.

An alternative to group punishments is to allow agents to make side payments to each other (*Goldlücke and Kranz* (2012, 2013)). This arrangement avoids costly forms of retaliation when an agent deviates, and yet is still incentive-compatible since the non-deviant agent receives a positive money transfer from the deviant. *Harrington and Skrzypacz* (2007, 2011) describe how the lysine and citric acid cartels successfully used these types of contracts, and employed monitors to audit the money-transfer process. This class of models offer a recursive characterization of the equilibrium set using simple penal codes, but is limited to teams of two agents or settings in which individual actions are observable.

More in general, our analysis is concerned with team production when a static game is repeated for infinitely many periods. In this setting, agents have an opportunity to retaliate against the team in future periods if shirking is detected (*Fudenberg and Maskin*, 1986; *Ostrom et al.*, 1992). Moreover, in repeated settings enforcing the aforementioned mechanisms of peer evaluations and relative performance rankings can become strategic problems in their own right, as exemplified by *Che and Yoo* (2001), *Fuchs* (2007), and *Cheng* (2016). Finally, our question bears some similarity to the "Who will guard the guardians?" question examined in *Hurwicz* (2008), *Rahman* (2012), *Aldashev and Zanarone* (2017), and *Acemoglu and Wolitzky* (2015) among others. Our setup differs slightly in that the guardian is the team itself, and individual team members must be willing to retaliate against the team when

group punishments are not enforced.

3.2 A Generalized Model of Repeated Team Production

We begin by describing a model of repeated team production where a benevolent Principal can impose group punishments after observing aggregate deviations. We provide conditions under which the Principal's ability to impose static group punishments—defined as punishments that occur *after* aggregate output is observed, but *before* currrent-period payoffs are realized—can be sustained in equilibrium to increase the welfare of the team. Moreover, if team members are sufficiently patient, the threat of these punishments can strictly increase team welfare relative to an environment where the Principal's actions are restricted to never implement group punishments.

3.2.1 Stage Game

A team consists of n *agents* indexed by $i = 1, \ldots, n$.[2] Each agent chooses an unobservable and nonnegative action $a_i \in \mathbb{R}_+$, representing a level of effort. The cost of action a_i is given by $c(a_i)$, where $c'(a_i) > 0$, $c''(a_i) \geq 0$, and $c(0) = 0$. Moreover, we write

$$a_{-i} = (a_1, \ldots, a_{i-1}, a_{i+1}, \ldots, a_n), a = (a_i, a_{-i}),$$

where the vector a constitutes an *effort profile*. An effort profile determines the aggregate outcome of team production according to a generic outcome function $x : \mathbb{R}_+^n \to \mathbb{R}_+$.

In addition to team members, a benevolent Principal (a construct for team payoffs) observes the aggregate outcome x and chooses a group punishment $\tau \geq 0$ that reduces the team's aggregate outcome. A strategy for the Principal is therefore $\tau : \mathbb{R}_+ \to \mathbb{R}_+$. For notational convenience, we define the final result of the team's effort after the Principal imposes punishments as the *aggregate net outcome function* $\ell(a, \tau)$, where $\ell : \mathbb{R}_+^{n+1} \to \mathbb{R}_+$.

[2] In what follows, we use the terms "agents"and "team members" interchangeably.

We make two sets of assumptions on this aggregate net outcome function. First, $\ell_\tau(a,\tau) < 0$, where the subscript denotes the partial derivative of $\ell(\cdot)$ with respect to τ. This assumption reflects the fact that in our model the Principal is just a construct for the team. Since the only resource available to the Principal is the outcome of team production, the Principal can never increase this outcome using group punishments. In other words, our first assumption rules out external subsidies from the model. Second, to keep the analysis close to *Holmström* (1982) we assume that for all i, j, $\ell_{a_i}(a,\tau) = \ell_{a_j}(a,\tau) \geq 0$ and $\ell_{a_i a_j}(a,\tau) \leq 0$, where the subscripts again denote partial derivatives.

Finally, the net outcome ℓ is distributed among team members according to a predetermined set of sharing rules $\{s_i\}_{i=1}^n$, where each $s_i \in (0,1)$ and

$$\sum_{i=1}^{n} s_i = 1. \tag{3.1}$$

To keep our analysis concise, we limit ourselves to cases where $s_i = 1/n$. This assumption can be relaxed to other sharing rules as long as each s_i is constant and (3.1) is satisfied.

Team members have identical preferences over their share of the aggregate outcome. Utility is given by $\pi : \mathbb{R}_+ \to \mathbb{R}$ which satisfies standard assumptions $\pi'(\ell) > 0$, $\pi''(\ell) \leq 0$, and $\lim_{\ell \to 0} \pi(\ell) = -\infty$. Additionally, utility from output interacts with individual effort according to a function $f : \mathbb{R}_+ \to \mathbb{R}_+$, which satisfies $f'(a_i) \leq 0$ and $f''(a_i) \leq 0$. The function $f(\cdot)$ represents possible interactions between the common payoff component, $\ell(a,\tau)$, and the individual agent's private effort a_i, and its interaction with $\pi(\cdot)$ allows us to nest the *Abreu* (1986) repeated oligopoly model within our generalized framework. In the oligopoly model, $\pi(\cdot)$ and $f(\cdot)$ respectively correspond to prices and quantities. Prices, like output shares, are common across all agents. Quantities, however, can vary across agents.[3] In our more general setting, one interpretation sees $f(\cdot)$ as part of a labor/leisure trade-off, while the cost function $c(\cdot)$ reflects all other personal costs related to production. The important feature that $f(\cdot)$

[3]The fact that oligopoly prices decrease in q while output shares increase in a is offset by $f(a)$ decreasing in a while q is increasing (in itself).

captures is that private and public gains from effort have a nontrivial interaction. In this general model, we can discipline this interaction more explicitly through our assumptions on $f(\cdot)$. Later on, we remove this interaction and find that a principal has no ability to improve outcomes.[4]

Since the Principal ignores sunk effort costs $c(\cdot)$, payoffs to the agents and Principal are respectively

$$u(a_i, a_{-i}, \tau) = \pi(s_i \ell(a_i, a_{-i}, \tau)) f(a_i) - c(a_i), \tag{3.2}$$

$$w(a, \tau) = \sum_{i=1}^{n} \pi(s_i \ell(a, \tau)) f(a_i). \tag{3.3}$$

3.2.1.1 Stage Game Equilibrium

A symmetric *perfect-public equilibrium* of the stage game consists of effort choices a_i by team members and a group punishment choice $\tau(x)$ by the Principal such that for every x, $\tau(x)$ maximizes (3.3) and such that given τ and a_{-i}, a_i maximizes (3.2)

Since in a static setting it is optimal for the principal not to impose group punishments (i.e. to set $\tau(x) = 0$), the optimal effort a^N of the static equilibrium, which we denote by a^N, is given by

$$a_i^N = \text{argmax}_{a_i} \left[\pi(s_i \ell(a_i, a_{-i}^N, 0)) f(a_i) - c(a_i) \right]. \tag{3.4}$$

Note that facing the Principal's optimal decision not to impose group punishments, the socially-optimal level of effort a^* which maximizes the sum of individual utilities is given by

$$a^* = \text{argmax}_a \sum_{i=1}^{n} u(a_i, a_{-i}, 0). \tag{3.5}$$

[4]The assumption that $\lim_{\ell \to 0} \pi(\ell) = -\infty$ is only needed when $\ell_{a_i} \geq 0$ to ensure that the team members can impose unbounded punishments on each other. On the other hand, the assumption that $f'(\cdot) \leq 0$ is necessary to guarantee the problem has an interior solution when $\ell_{a_i} \geq 0$. More generally, the necessary assumption for the repeated model of team production to have an interior solution is that $\text{sign}(\ell_{a_i}) = -\text{sign}(f')$. The assumption that $f''(\cdot) \leq 0$ is sufficient but not necessary to obtain our results, and allows us to easily compare the generalized model with the repeated oligopoly model of *Abreu* (1986) in Section 3.3.

In the following Lemma 3.2.1.1, we establish that the equilibrium level of effort of this static game is smaller than the socially-optimal level of effort.

Lemma III.1. $0 < a_i^N < a_i^*$.

Proof. An individual agent's first-order conditions yield

$$s_i \ell_{a_i}(a_i, a_{-i}, 0)\pi'(s_i \ell(a_i, a_{-i}, 0))f(a_i) + f'(a_i)\pi(s_i \ell(a_i, a_{-i}, 0)) \;=\; c'(a_i). \tag{3.6}$$

The profile a^N necessarily satisfies (3.6) for all agents $i = 1, \ldots, n$. That is,

$$s_i \ell_{a_i}(a^N, 0)\pi'(s_i \ell(a^N, 0))f(a_i^N) + f'(a_i^N)\pi(s_i \ell(a^N, 0)) \;=\; c'(a_i^N). \tag{3.7}$$

The first order condition for the socially-optimal level of effort, on the other hand, implies that for all i

$$\begin{aligned} c'(a_i^*) \;=\;& s_i \ell_a(a^*, 0)\pi'(s_i \ell(a^*, 0))f(a_i^*) + f'(a_i^*)\pi(s_i \ell(a^*, 0)) \\ &+ \sum_{j \neq i} s_j \ell_{a_i}(a^*, 0)\pi'(s_j \ell(a^*, 0))f(a_j^*). \end{aligned} \tag{3.8}$$

Conditions (3.6) and (3.8) differ by an additional term in (3.8). This extra term represents the positive externality of one agent's additional effort on the remaining $(n-1)$ agents. Since $\pi' > 0$, $s_i \in [0, 1]$, and $f(a_j) > 0$ for any $a_j > 0$, the additional term is necessarily positive. This implies that

$$\begin{aligned} &s_i \ell_{a_i}(a^N, 0)\pi'(s_i \ell(a^N, 0))f(a_i^N) + f'(a_i^N)\pi(s_i \ell(a^N, 0)) - c'(a_i^N) \;> \\ &\quad s_i \ell_{a_i}(a^*, 0)\pi'(s_i \ell(a^*, 0))f(a_i^*) + f'(a_i^*)\pi(s_i \ell(a^*, 0)) - c'(a_i^*). \end{aligned} \tag{3.9}$$

The result follows from our assumptions on $\ell(\cdot)$, $\pi(\cdot)$, and $f(\cdot)$. Since $\lim_{\ell \to 0^+} \pi(\ell) = -\infty$, we rule out the boundary solution $a_i^N = 0$, so $0 < a_i^N < a_i^*$. □

Note that if the Principal were able to commit to group punishments when the aggregate outcome is smaller than $x(a^*)$, then each producer contributing a_i^* would be an equilibrium. For example, for a given effort profile a, if the Principal's strategy was to implement some $\tau(x(a)) > 0$ such that $\ell(a, \tau(x(a))) = 0$ if $x(a) < x(a^*)$, and conversely to implement $\tau = 0$ if $x(a) = x(a^*)$, then each agent's best response to a_{-i}^* would be to choose $a_i = a_i^*$.[5] In this sense, the threat of group punishments would be useful if the Principal could commit to such a strategy. In the next section, we investigate whether group punishments may be sustainable and welfare-improving when agents and the Principal interact repeatedly. Before proceeding to the repeated game, we establish the intermediate result that agents will increase their effort in the interior of $[a_i^N, a_i^*]$ when $a_{-i} < a_i^N$.

Corollary III.2. *If $a_{-i} < a_i^N$, then the most profitable deviation a_i' is such that $a_i' > a_i^N$.*

Proof. Consider the condition that is satisfied when $a_i = a_i^N$ for $i = 1, \ldots, n$.

$$s_i \ell_{a_i}(a^N, 0)\pi'(s_i \ell(a^N, 0)) f(a_i^N) + f'(a_i^N)\pi(s_i \ell(a^N, 0)) \;=\; c'(a_i^N). \tag{3.10}$$

Now suppose that the effort by all other producers but i (denoted by a_{-i}) decreases from a_i^N. Then,

$$s_i \ell_{a_i}(a_i^N, a_{-i}, 0)\pi'(s_i \ell(a_i^N, a_{-i}, 0)) f(a_i^N) + f'(a_i^N)\pi(s_i \ell(a_i^N, a_{-i}, 0)) \;>\; c'(a_i^N). \tag{3.11}$$

The optimal response a_i' by agent must satisfy the first-order condition

$$s_i \ell_{a_i}(a_i', a_{-i}, 0)\pi'(s_i \ell(a_i', a_{-i}, 0)) f(a_i') + f'(a_i')\pi(s_i \ell(a_i', a_{-i}, 0)) \;=\; c'(a_i'), \tag{3.12}$$

which means that the right-hand side of (3.11) must increase and/or its left-hand side must decrease. Therefore, $a_i' > a_i^N$. □

[5] In this example, we assume that for each a, there always exists some $\tau(x(a)) > 0$ such that $\ell(a, \tau(x(a))) = 0$. In other words, we assume that there exists a punishment such that the Principal can completely destroy the aggregate outcome.

3.2.2 Infinitely-Repeated Game

In this section, we develop and analyze an infinitely-repeated version of the static team production model described above. We focus on symmetric, perfect-public equilibria and illustrate how team members may incentivize the Principal such that group punishments are sustainable in equilibrium even when the Principal lacks commitment. We go on to show that along the best equilibrium path, group punishments are not implemented. However, the threat of group punishments allows team members to attain strictly higher welfare than they would in an economy where group punishments are not allowed—the Principal's actions are restricted to never impose group punishments.

3.2.2.1 Histories, Perfect-Public Equilibria, and One-Shot Deviations

Here we describe the infinitely-repeated game, define our notion of equilibrium, and simplify our equilibrium characterization by appealing to the one-shot deviation principle. Proposition III.3 of this section shows that the entire set of perfect-public equilibria can be attained by preventing single-period (one-shot) deviations in the infinitely-repeated game.

Let $h_t^w \in H^w$ where $H^w = \mathbb{R}_+^2$ denote the public outcomes (x_t, τ_t) observed at the end of period t. Then, let $\mathcal{H}^w$ denote set of public histories with $\mathcal{H}^w = \bigcup_{t=0}^{\infty} (H^w)^t$. Similarly, define the set of histories for agent i as $\mathcal{H}^i = \bigcup_{t=0}^{\infty} (\mathbb{R}_+ \times H^w)^t$. A pure strategy for agent i is a mapping from the set of all possible agent i histories into the set of pure actions,

$$\sigma_i : \mathcal{H}^i \to \mathbb{R}_+.$$

A pure strategy for the Principal is a mapping from the set of public histories and an observation of the aggregate outcome into the set of pure actions for the Principal,

$$\sigma_w : \mathcal{H}^w \times \mathbb{R}_+ \to \mathbb{R}_+.$$

We assume agents and the Principal have a common discount factor δ and restrict attention to public strategies which are functions only of the public history. Given a strategy profile $\sigma = (\{\sigma_i\}_{i=1}^n, \sigma_w)$, if $h^{wt} \in H^{wt}$ denotes a generic period-t history, we let $U_i^t(h^{wt}, \sigma)$ denote the discounted continuation payoffs agent i obtains from period t onwards. Since the Principal chooses an action after period-t effort decisions are sunk, the Principal's discounted continuation payoffs satisfy

$$U_t^w\left(h^{wt}, \sigma\right) = \sum_i U_t^i\left(h^{wt}, \sigma\right) + (1-\delta)\, c \sum_i \sigma_i\left(h^{wt}\right). \tag{3.13}$$

In Appendix C.2.1.1 we define continuation games and strategies, perfect-public equilibria, and one-shot deviations. In the next proposition, we prove that equilibria can be constructed recursively by ensuring that for any history, neither the agents nor the Principal have a profitable one-shot deviation.

Proposition III.3. *A strategy profile $\sigma = (\{\sigma_i\}_{i=1}^n, \sigma_w)$ is perfect-public if and only if there are no profitable one-shot deviations for the agents and there are no profitable one-shot deviations for the Principal.*

Proof. See Appendix C.2.1.2. □

3.2.2.2 Equilibrium Set Characterization

We now describe a procedure to characterize the set of symmetric equilibrium payoffs using carrot-and-stick strategies as in *Abreu* (1986). As we will argue, individual deviations by team members may be subject to group punishments chosen by the Principal. However, limited commitment of the Principal implies that agents will need to impose discipline on the Principal in the event that the Principal attempts to avoid the static losses associated with group punishments. Nonetheless, we will show that extremal equilibrium payoffs (both the best and the worst equilibrium payoff) need not feature group punishments.

We focus on characterizing *strongly* symmetric equilibria, and we therefore simplify our

notation by dropping i subscripts and by using a in place of $(a, a, \ldots, a)$ for producers' strategies, $u(a, 0)$ in place of $u_i(a, a, \ldots, a, \tau = 0)$ for producers' payoffs and so on.

Under the one-shot deviation principle, given the worst perfect-public equilibrium payoff $\underline{v}$, the best perfect-public equilibrium payoff $\bar{v}$ can be constructed as the solution to the following program:

$$\bar{v} = \max_{a, \tau(\cdot), v(\cdot, a, \tau(\cdot))} u(a, 0), \tag{3.14}$$

subject to, for all a'

$$u(a, 0) \geq (1 - \delta) u(a', a, \tau(x(a', a))) + \delta v(a', a, \tau(x(a', a))) \tag{3.15}$$

$$v(a', a, \tau(x(a', a))) \in [\underline{v}, \bar{v}], \tag{3.16}$$

and

$$(1 - \delta) w(a', a, \tau(x(a', a))) + n\delta v(a', a, \tau(x(a', a))) \geq (1 - \delta) w(a', a, 0) + n\delta \underline{v}. \tag{3.17}$$

Inequality (3.15) represents the incentive compatibility constraint for each agent, which requires the symmetric payoff $u(a, 0)$ to be greater or equal to the payoff associated with a deviation effort a' with static payoff $u(a', a, \tau(x(a', a)))$ and continuation payoff $v(a', a, \tau(x(a', a))$. Equation (3.16) represents the feasibility constraint for the continuation payoff $v(a', a, \tau(x(a', a))$, which must lie between the worst equilibrium payoff $\underline{v}$ and the best equilibrium payoff $\bar{v}$. Finally, (3.17) is the incentive compatibility constraint for the Principal, requiring the Principal to have sufficient incentives to enforce the prescribed group punishment once one of the n team members deviates to a'. The left-hand side of (3.17) is the Principal's payoff when implementing the prescribed group punishment while the right-hand side is the payoff from a deviation to $\tau = 0$, followed by the worst perfect-public equilibrium payoff $\underline{v}$.

It is useful here to reduce the dimensionality of the problem by eliminating the Principal's

incentive-compatibility constraint. Since (3.17) must bind in any solution to the above program, the continuation payoff following a deviation by an agent must satisfy

$$v\left(a', a, \tau\left(x\left(a', a\right)\right)\right) = \underline{v} + \frac{1-\delta}{\delta}\frac{1}{n}\left[w\left(a', a, 0\right) - w\left(a', a, \tau\left(x\left(a', a\right)\right)\right)\right]. \tag{3.18}$$

Hence, for any deviation a', we may write the agent's incentive-compatibility constraint (3.15) as

$$u\left(a', a, 0\right) \geq (1-\delta)\left[u\left(a', a, \tau\left(x\left(a', a\right)\right)\right) + \frac{1}{n}\left[w\left(a', a, 0\right) - w\left(a', a, \tau\left(x\left(a', a\right)\right)\right)\right]\right] + \delta\underline{v}. \tag{3.19}$$

Let $g\left(a', a, \tau\left(x\left(a', a\right)\right)\right)$ denote the static payoff for an individual agent exerting effort a' when all other producers producers produce a—the term in the outer square brackets on the right-hand side of (3.19). We call this quantity the *total static deviation payoff*. Using this definition, we re-write the problem (3.14)-(3.17) as

$$\bar{v} = \max_{a} u\left(a, 0\right), \tag{3.20}$$

subject to, for all a',

$$u\left(a, 0\right) \geq (1-\delta)\, g\left(a', a, \tau\left(x(a', a)\right)\right) + \delta\underline{v}, \tag{3.21}$$

$$\bar{v} \geq \frac{1-\delta}{\delta}\frac{1}{n}\left[w\left(a', a, 0\right) - w\left(a', a, \tau\left(x(a', a)\right)\right)\right] + \underline{v}, \tag{3.22}$$

$$g\left(a', a, \tau\left(x(a', a)\right)\right) = u\left(a', a, \tau\left(x(a', a)\right)\right) - \frac{1}{n}\left[w\left(a', a, \tau\left(x(a', a)\right)\right) - w\left(a', a, 0\right)\right]. \tag{3.23}$$

Next, it is useful to define the maximum deviation payoff an agent can achieve by devi-

ating to a' from profile a, which we denote by $\hat{g}\left(a, \tau\left(\cdot\right)\right)$. This payoff satisfies

$$\hat{g}\left(a, \tau\left(\cdot\right)\right) = \max_{a'} g\left(a', a, \tau(x(a', a)\right).$$

In the next lemma, we show that as long as the prescribed level of effort is smaller than the static Nash equilibrium level of effort, the maximum deviation payoff $\hat{g}(a, \tau\left(\cdot\right))$ is minimized when the Principal imposes no group punishments (i.e. when $\tau = 0$).

Lemma III.4. *Supppose that $f'(a) < 0$. Then $\hat{g}\left(a, \tau\left(\cdot\right)\right) \geq \hat{g}\left(a, \tau = 0\right)$ when $a \leq a^N$.*

Proof. For notational simplicity, we remove the dependency of $\tau(\cdot)$ on its arguments. Note that

$$\begin{aligned}
\frac{\partial g}{\partial \tau} &= s_i \ell_\tau(a', a, \tau)\pi'(s_i \ell(a', a, \tau)) f(a') \\
&- \frac{1}{n} s_i \ell_\tau(a', a, \tau)\pi'(s_i h(a', a, , \tau))\left[(n-1)f(a) + f(a')\right] \qquad (3.24) \\
&= s_i \ell_\tau(a', a, \tau)\pi'(s_i \ell(a', a, \tau))\frac{n-1}{n}\left[f(a') - f(a)\right]. \qquad (3.25)
\end{aligned}$$

Since $\ell_\tau \leq 0$ and $\pi' > 0$, for $\partial g / \partial \tau > 0$ we need only show that $[f(a') - f(a)] < 0$. This is true, however, since $f(a)$ is decreasing and $a' > a$ by Corollary III.2. □

Lemma III.4 establishes that group punishments ($\tau\left(\cdot\right) > 0$) increase the incentives of individual agents to deviate when $a \leq a^N$. Intuitively, when the perscribed level of effort is smaller than the static Nash equilibrium level, each agent has a (static) incentive to exert more effort. Imposing group punishments for excess effort in this region simply strengthens individual agents' incentives to exert effort, and therefore has no use in enforcing the prescribed behavior.

Lemma III.4 plays a key role in allowing us to characterize simple equilibrium strategies which obtain the the infimum perfect-public equilibrium payoff $\underline{v}$. To construct $\underline{v}$, we propose a carrot-and-stick strategy, which with a small abuse of notation we write as $\sigma\left(\left(\tilde{a}, \bar{a}\right), (0, 0)\right)$. This strategy calls for agents to play some "stick" level of effort $\tilde{a}$ and subsequently revert to

the "carrot" level $\bar{a}$—the level of effort prescribed in the best perfect-public equilibrium. If either the carrot or the stick are played by all agents as prescribed by the strategy, the Principal chooses $\tau = 0$. If the Principal detects an aggregate deviation $x(a', \bar{a}) \neq x(\bar{a})$ from the carrot $\bar{a}$, the Principal chooses to implement a group punishment $\tau(x(a', \bar{a})) > 0$, and the agents consequently revert to some strategy with value $v(a', \bar{a}, \tau(x(a', \bar{a}))$. If the Principal observes an aggregate deviation $x(a', \tilde{a}) \neq x(\tilde{a})$ from the stick $\tilde{a}$, the Principal chooses $\tau(x(a', \tilde{a})) = 0$, and the producers consequently revert to the carrot-and-stick strategy $\sigma\left((\tilde{a}, \bar{a}), (0, 0)\right)$ with value $\underline{v}$. Finally, any deviation by the Principal causes the carrot-and-stick strategy to be repeated.

Proposition III.5. *There exists an output $\tilde{a}$ such that the carrot-and-stick strategy $\sigma\left((\tilde{a}, \bar{a}), (0, 0)\right)$ attains the value $\underline{v}$—that is, $\sigma\left((\tilde{a}, \bar{a}), (0, 0)\right)$ is an optimal punishment.*

Proof. Given $\underline{v}$, the infimum of symmetric perfect-public equilibrium payoffs and hence $\bar{a}$ (the value that attains the maximum, $\bar{v}$ in the program (3.20)-(3.23)), we may obtain $\tilde{a}$ such that

$$\underline{v} = (1 - \delta)\, u\left(\tilde{a}, 0\right) + \delta u\left(\bar{a}, 0\right). \tag{3.26}$$

We now argue that the carrot-and-stick strategy $\sigma\left((\tilde{a}, \bar{a}), (0, 0)\right)$ is an equilibrium. By construction, the punishment has value $\underline{v}$. Since deviations from $\bar{a}$ are unprofitable when punished by $\underline{v}$, they are by construction unprofitable when punished by $\sigma\left((\tilde{a}, \bar{a}), (0, 0)\right)$.

To show that no producer wishes to deviate when prescribed to contribute effort $\tilde{a}$, we must show that for all a',

$$\underline{v} = (1 - \delta)\, u\left(\tilde{a}, 0\right) + \delta u\left(\bar{a}, 0\right) \quge (1 - \delta)\, g\left(a', \tilde{a}, 0\right) + \delta \underline{v}, \tag{3.27}$$

and in particular

$$\underline{v} = (1 - \delta)\, u\left(\tilde{a}, 0\right) + \delta u\left(\bar{a}, 0\right) \quge (1 - \delta)\, \hat{g}\left(\tilde{a}, 0\right) + \delta \underline{v}. \tag{3.28}$$

We proceed by contradiction. Suppose (3.28) does not hold. Then there must exist another (strongly symmetric) equilibrium σ^y with first-period output $a^y \leq a^N$ such that

$$(1-\delta)\,\hat{g}\,(\tilde{a},0) + \delta\underline{v} > (1-\delta)\,u\,(a^y,0) + \delta U\,(\sigma^y|_{a^y}) \geq \underline{v} \tag{3.29}$$

where $U\,(\sigma^y|_{a^y})$ is the continuation payoff to a single producer from σ^y after contributing a^y in the first period.[6]

Replacing the definition of $\underline{v}$ in (3.29) implies

$$(1-\delta)\,u\,(a^y,0) + \delta U\,(\sigma^y|_{a^y}) \quad \geq \quad (1-\delta)\,u\,(\tilde{a},0) + \delta u\,(\bar{a},0)\,. \tag{3.30}$$

Since $U\,(\sigma^y|_{a^y}) \leq u\,(\bar{a},0)$, it must be that $u\,(a^y,0) \geq u\,(\tilde{a},0)$ and therefore $a^y \geq \tilde{a}$. However, we will show that if σ^y is a perfect-public equilibrium, $\tilde{a} > a^y$, yielding the necessary contradiction. Since σ^y is an equilibrium,

$$(1-\delta)\,u\,(a^y,0) + \delta U\,(\sigma^y|_{a^y}) \quad \geq \quad (1-\delta)\,\hat{g}\,(a^y,\tau\,(x(a^y))) + \delta\underline{v}, \tag{3.31}$$

so that from (3.29)

$$(1-\delta)\,g\,(\tilde{a},0) + \delta\underline{v} \quad > \quad (1-\delta)\,\hat{g}\,(a^y,\tau\,(x(a^y))) + \delta\underline{v}. \tag{3.32}$$

Since $a^y \leq a^N$, Lemma III.4 implies that

$$\hat{g}\,(a^y,\tau\,(y(a^y))) \geq \hat{g}\,(a^y,0) \tag{3.33}$$

[6]Since repeated play of the static Nash equilibrium output a^N with no punishments must be an equilibrium, it is straightforward to show that the prescribed effort under the "stick" must satisfy $\tilde{a} \leq a^N$. If $a^y > a^N$, however, (3.29) implies that

$$\hat{g}(\tilde{a},0) > \hat{g}(a^N,0).$$

Since the best deviation payoff in the absence of punishments is increasing in a, this would imply $a^N < \tilde{a}$, a contradiction.

so that

$$\hat{g}\left(\tilde{a},0\right) > \hat{g}\left(a^{y},0\right). \tag{3.34}$$

Since $\hat{g}\left(a,0\right)$ is increasing in a, (3.34) implies $\tilde{a} > a^{y}$ providing the needed contradiction. □

Proposition III.5 greatly simplifies the characterization of the set of perfect-public equilibrium payoffs. We have shown that the worst equilibrium payoff can be attained without requiring group punishments (either on the equilibrium path, or off the equilibrium path following deviations from the "stick"). The key feature of our economy which yields this result is the fact that during the "stick" phase of the worst equilibrium strategy, group punishments actually make deviations from the stick more appealing to producers. Consequently the optimal strategy for the Principal is to not impose group punishments. Using the results from Proposition III.5, we now characterize strategies that allow us to attain the entire set of perfect-public equilibria.

Proposition III.6. *If the strategy σ is a Perfect-Public Equilibrium, then $u\left(\sigma\right) \in [\underline{v}, \bar{v}]$. If $v \in [\underline{v}, \bar{v}]$, then there exists a Perfect-Public Equilibrium strategy σ such that $u\left(\sigma\right) = v$.*

Here we provide a sketch of the argument and leave a formal proof to Appendix C.2.1.3. It is clear that any equilibrium satisfies the constraints of the program (3.14)-(3.17) and therefore $U(\sigma) \in [\underline{v}, \bar{v}]$. It only remains to show that any value in this set may be attained by some equilibrium strategy. We prove this result using an induction argument. To begin, it is straightforward to characterize the set of values that can be attained with strategies which restrict the Principal never to impose punishments (either on or off the equilibrium path). This set, which we denote $[\underline{v}^{A}, \bar{v}^{A}]$ defines the set of values that are attainable as subgame-perfect equilibria, and can be easily constructed with carrot-and-stick strategies following *Abreu* (1986).[7]

[7]We use subgame-perfect equilibria as our benchmark, as opposed to renegotiation-proof equilibria. The characterization of renegotiation-proof equilibria in repeated games is complex and can depend on the model's

Since $\underline{v}^A < \bar{v}^A$, it is feasible to sustain one period of punishments in the event some agent deviates from a prescribed level of effort. We therefore construct equilibria in which all agents are asked to contribute some effort level a. If all agents do so, then no punishments are implemented and the strategy repeats. If some agent deviates to some a'—so that the aggregate outcome is different than $x(a)$—then the Principal is called upon to implement a punishment. If the Principal implements the prescribed punishment, agents play some equilibrium without punishments which delivers the value $v(a', a, \tau)$. If the Principal does not implement the prescribed punishment, agents play the strategy associated with the worst equilibrium of a model where punishments are not allowed, with value $\underline{v}^A$. We choose a positive but sufficiently small punishment τ to ensure that $v(a', a, \tau) \in \left(\underline{v}^A, \bar{v}^A\right]$ for all relevant deviations a'. We show that this strategy delivers equilibrium values $u(a) > \bar{v}^A$. Given these strategies, we are able to construct carrot-and-stick equilibrium strategies which deliver values strictly below $\underline{v}^A$. In following these steps, we have constructed an operator which maps equilibrium value sets supported by perfect-public equilibrium strategies into similar sets that are strictly larger and yet still attainable with perfect-public equilibrium strategies. We show that repeated application of this operator starting from a set where group punishments are not part of the Principal's action set necessarily converges to the set $[\underline{v}, \bar{v}]$ defined by the program (3.14)-(3.17). In this way, we construct a perfect-public equilibrium strategy which delivers each value $v \in [\underline{v}, \bar{v}]$.

We now use Proposition III.7 to fully characterize the values of the best and worst perfect-public equilibrium payoffs.

Proposition III.7. *The optimal carrot-and-stick punishment satisfies*

$$\hat{g}(\tilde{a}, 0) = (1-\delta) u(\tilde{a}, 0) + \delta u(\bar{a}, 0) = \underline{v}, \tag{3.35}$$

$$\hat{g}(\bar{a}, \tau(\cdot)) = u(\bar{a}, 0) + \delta (u(\bar{a}, 0) - u(\tilde{a}, 0)) \text{ if } \bar{a} < a^*, \tag{3.36}$$

parameters (see, e.g. *Aramendía et al.* (2005)), which makes the renegotiation-proof set a less amenable benchmark for our model. Here we want to emphasize that our setup has the flavor of within-period renegotiation, excluding the possibility of renegotiation in future periods.

and

$$\hat{g}\left(\bar{a},\tau\left(\cdot\right)\right) \quad \leq \quad u\left(\bar{a},0\right)+\delta\left(u\left(\bar{a},0\right)-u\left(\tilde{a},0\right)\right) \;\; \textit{if } \bar{a}=a^{*}. \tag{3.37}$$

The proof is a straightforward extension of those found in *Abreu* (1986) and hence relegated to the Appendix (see Section C.2.1.4). Propositions III.5 and III.7 show that neither the best nor the worst perfect-public equilibria feature group punishments imposed by the Principal. Nonetheless, we will show momentarily that the out-of-equilibrium threat of group punishments allows team members to attain higher welfare than in an economy where group punishments are not part of the Principal's action set. For expositional brevity, we will refer to such economy as an economy where group punishments "are not allowed." Let $\bar{a}^{A}$ and $\tilde{a}^{A}$ respectively denote the carrot and stick levels of output in the model where group punishments are not allowed. Similarly, let $\bar{v}^{A}$ and $\underline{v}^{A}$ denote the best and worst perfect-public equilibrium values in the model where group punishments are not allowed. Proposition III.8 formally establishes that if the equilibrium output level $\bar{a}$ is sustained by a positive punishment threat (a deviation by an agent is followed by a strictly positive group punishment implemented by the Principal), then the presence of such a threat *strictly* improves welfare, or $\bar{v}>\bar{v}^{A}$.

Proposition III.8. *For any equilibrium output levels $\bar{a}\leq a^{*}$, $\bar{a}^{A}<\bar{a}$ if $\bar{a}$ is sustained by a positive punishment threat (for some $a'\neq\bar{a}$, $\tau\left(x\left(a',\bar{a}\right)\right)>0$), then $\bar{v}=u\left(\bar{a},0\right)>u\left(\bar{a}^{A},0\right)=\bar{v}^{A}$.*

Proof. First, note that since the Principal can always choose $\tau=0$, $\left[\underline{v}^{A},\bar{v}^{A}\right]\subseteq\left[\underline{v},\bar{v}\right]$. Therefore $u\left(\bar{a},0\right)\geq u\left(\bar{a}^{A},0\right)$, or $\bar{a}\geq\bar{a}^{A}$. Now suppose by contradiction that if $\bar{a}$ is sustained by a positive threat $\tau>0$, then $\bar{a}=\bar{a}^{A}$. Since $\bar{a}=\bar{a}^{A}>a^{N}$, $\hat{g}\left(\bar{a}^{A},0\right)=\hat{g}\left(\bar{a},0\right)>\hat{g}\left(\bar{a},\tau\right)$. From (3.36),

$$u\left(\bar{a}^{A},0\right)+\delta\left(u\left(\bar{a}^{A},0\right)-u\left(\tilde{a}^{A},0\right)\right) \quad > \quad u\left(\bar{a},0\right)+\delta\left(u\left(\bar{a},0\right)-u\left(\tilde{a},0\right)\right), \tag{3.38}$$

or

$$u\left(\tilde{a},0\right) \;>\; u\left(\tilde{a}^{A},0\right). \tag{3.39}$$

But from (3.35), this implies

$$\underline{v} = (1-\delta)\,u\left(\tilde{a},0\right) + \delta u\left(\bar{a},0\right) \;>\; (1-\delta)\,u\left(\tilde{a}^{A},0\right) + \delta u\left(\bar{a}^{A},0\right) = \underline{v}^{A}, \tag{3.40}$$

a contradiction with $\left[\underline{v}^{A},\bar{v}^{A}\right] \subseteq [\underline{v},\bar{v}]$. □

We conclude this section by providing conditions on agents' static payoffs such that group punishments improve welfare. Specifically, we note that the assumption underlying our Lemma III.4 and Propositions III.5 to III.8 is that the private utility component $f(a_i)$, is decreasing in effort. Proposition III.9 considers the alternative case where $f(a_i)$ is constant. We find that the interaction between private and publicly observed payoffs is essential in enabling static group punishments to enlarge the equilibrium set, relative to an economy where punishments are not allowed.

Proposition III.9. *Let κ be some constant. If $f(a) = \kappa$, for all $a \in [0, a^*]$, static group punishments do not improve equilibrium outcomes relative to a model where the Principal is not allowed to impose group punishments.*

Proof. This result is clear from Equation (3.25). If $f(a) = f(a') = \kappa$, then $\partial g/\partial \tau = 0$ and group punishment have no effect on producers' payoffs. □

Proposition III.9 states that a necessary condition for static group punishments to improve welfare is the presence of complementarities between aggregate outcomes and private actions in the individual agents' stage game payoffs. Absent these complementarities (i.e. when $f(a) = \kappa$), group punishments have no effect on the total deviation payoff g because the impact of the punishment on the static deviation gain $u\left(a', a, \tau\left(x(a', a)\right)\right)$ is the exactly

equal to the impact that these punishments have on the per-capita share of the cost to incentivize the Principal, $[w(a', a, \tau(x(a', a))) - w(a', a, 0)]/n$.

When team members' private actions instead interact with aggregate outcomes (i.e. $f(a)$ is not constant in a), then group punishments can reduce team members' private incentives to deviate through the interaction of these private incentives with the aggregate outcome. In these cases, group punishments are useful to deter individual deviations, and an outsider is not needed to improve welfare. In other words, in presence of complementarities between aggregate and individual outcomes, the team (represented by the Principal) can implement budget-breaking static punishments that improve welfare without requiring the intervention of an outsider.

3.3 An Application: Repeated Oligopoly with a Principal

In this section, we apply our generalized team production model to the repeated oligopoly model of *Abreu* (1986). We start by characterizing the stage game payoffs and equilibria, and we then provide a numerical illustration of our main result that group punishments increase team welfare in a repeated setting. In Section 3.3.3, we show how different degrees of interaction between oligopolistic producers can impact the effectiveness of group punishments.

3.3.1 Stage Game

A team is composed by n *producers* indexed by $i = 1, \ldots, n$. Each producer chooses an unobservable action $q_i \in \mathbb{R}_+$ where q_i represents a level of output generated by producer i. Each producer generates output at a constant marginal cost $c \in (0, 1)$. We let $q = (q_1, \ldots, q_n) \in \mathbb{R}^n_+$ and we write

$$q_{-i} = (q_1, \ldots, q_{i-1}, q_{i+1}, \ldots, q_n), q = (q_i, q_{-i}).$$

The producers' choices of output give rise to an aggregate quantity of output $Q = \sum_{i=1}^{n} q_i$. Each producer's stage-game strategy is simply $q_i \in \mathbb{R}_+^n$.

In addition to the producers, a benevolent Principal observes aggregate output Q and imposes an observable group punishment $\tau \in [0, 1]$, which represents an implicit tax imposed by the Principal on the consumers of the good. A strategy for the Principal is $\tau : \mathbb{R}_+ \to [0, 1]$.

The price at which producers sell their output is a function of aggregate output and the tax chosen by the Principal. Specifically,

$$p\left(Q, \tau\right) = \max\left\{(1 - \tau) - Q, 0\right\}. \tag{3.41}$$

This price function represents an inverse demand curve for consumers who face taxes τ on purchases of units of output. From (3.41) it is clear that the Principal's choice of the tax may reduce the price of output for all producers.

Given actions by the producers and the Principal, each producer's payoff is given by

$$u_i(q, \tau) = p\left(Q, \tau\right) q_i - c q_i. \tag{3.42}$$

We again assume that the Principal is benevolent in the sense that the Principal has preferences over a weighted average of the producers' utility. Since the Principal chooses the tax τ after production costs are sunk, the Principal's payoff from any level of total output Q and tax τ is given by

$$w(Q, \tau) = p\left(Q, \tau\right) Q. \tag{3.43}$$

Note that (3.42)-(3.43) immediately map to the generalized payoffs (3.2)-(3.3) when we i) impose symmetric sharing rules (i.e. $s_i = 1/n$), ii) impose linear utility, interaction and cost functions of the form $\pi(s_i \ell) = s_i \ell$, $f(a_i) = a_i$ and $c(a_i) = c a_i$, respectively, and iii) define the aggregate net outcome function as $\ell(a, \tau) = n \max\{1 - \tau - \sum_i a_i, 0\}$.[8]

[8]Contrary to our generalized model, the oligopoly model's net outcome function is such that, for all i, j, $\ell_{a_i}(a, \tau) = \ell_{a_j}(a, \tau) < 0$. This changes the sign of the main inequalities of our paper (for example, the

A symmetric *perfect-public equilibrium* in the stage game consists of choices for producers q_i and a Principal's strategy $\tau(Q)$ such that for every Q, $\tau(Q)$ maximizes (3.43) and given τ and q_{-i}, q_i maximizes (3.42). This equilibrium is straightforward to determine since for any Q, the Principal optimally chooses $\tau(Q) = 0$. Facing q_{-i} each producer's best response satisfies

$$q_i = \begin{cases} \frac{1}{2}\left(1 - \sum_{-i} q_{-i} - c\right) & \text{if } 1 - \sum_{-i} q_{-i} - c > 0, \\ 0 & \text{otherwise,} \end{cases} \tag{3.44}$$

with the equilibrium level of q_i satisfying

$$q_i^N = \frac{1-c}{n+1}. \tag{3.45}$$

Note that facing the Principal's optimal decision to set the tax equal to zero, the level of output which maximizes the producers' joint profits satisfies

$$q_i^m = \arg\max_{q_i} q_i\left(1 - nq_i - c\right), \tag{3.46}$$

with solution

$$q_i^m = \frac{1-c}{2n}. \tag{3.47}$$

From (3.45) and (3.47), observe that the level of output which maximizes joint producer profits is lower than the perfect-public equilibrium outcome. Intuitively, producer i has an incentive to generate more output when the other producers generate less than q_i^N and prices are high. In contrast, producer i has an incentive to generate less output when the other producers generate more than q^N.

Nash equilibrium level of output is larger than the socially-optimal level of output), but the procedure to characterize the set of equilibrium payoffs is identical to the procedure developed in the previous section.

3.3.2 Infinitely-Repeated Game

As in the previous sections, we focus on characterizing strongly symmetric equilibria. Following the same steps as in Section 3.2.2, it is easy to show that the generalized program (3.20)-(3.23) maps to the following program in the repeated oligopoly model:

$$\bar{v} = \max_{q} u(q, 0), \tag{3.48}$$

subject to, for all q',

$$u(q, 0) \geq (1-\delta)\, g(q', q, \tau(q' + (n-1)\, q)) + \delta \underline{v}, \tag{3.49}$$

$$\bar{v} \geq \frac{1-\delta}{\delta} \frac{1}{n} \left[w(q' + (n-1)\, q, 0) - w(q' + (n-1)\, q, \tau(q' + (n-1)\, q)) \right] + \underline{v}, \tag{3.50}$$

where $\underline{v}$ and $\bar{v}$ again denote the worst and the best perfect-public equilibrium payoffs of the repeated game, and where (using (3.42) and (3.43)) the total static deviation payoff $g(q', q, \tau(q' + (n-1)\, q))$ is given in closed-form by

$$\begin{aligned} g(q', q, \tau(q' + (n-1)\, q)) = q' \left((1 - \tau(q' + (n-1)\, q)) - (q' + (n-1)\, q) - c \right) \\ + \frac{1}{n} \tau(q' + (n-1)\, q)\, (q' + (n-1)\, q). \end{aligned} \tag{3.51}$$

This closed-form expression reveals that the static deviation payoff in the oligopoly model is comprised of two components. The first component can be re-written as $p(q', q, \tau(q' + (n-1)q))q'$, and represents the static payoff that the producer obtains by deviating to q' from q when the deviation is punished by a tax $\tau(q' + (n-1)q)$. The second component, $\tau(q' + (n-1)q)(q' + (n-1)q)/n$, is a cost that individual producers have to pay to incentivize the Principal to implement the prescribed punishment $\tau = \tau(q' + (n-1)\, q)$ as opposed to her most profitable deviation $\tau = 0$.

Finally, let $\hat{g}(q, \tau(\cdot))$ denote the maximum deviation payoff one producer can achieve from a deviation to q' when other producers generate q. As in Lemma III.4, we now show that as long as the prescribed output is larger than the static Nash equilibrium output, the maximum deviation payoff $\hat{g}(q, \tau(\cdot))$ is minimized when the Principal levies no taxes (i.e., when $\tau = 0$).

Lemma III.10. *$\hat{g}(q, \tau(\cdot)) \geq \hat{g}(q, \tau = 0)$ when $q \geq q^N$.*

Proof. See Appendix C.2.1.5. □

Using Lemma III.10, the results from Propositions III.5 to III.8 naturally extend to the repeated oligopoly model, and are therefore omitted for the sake of brevity. In particular, we find that the worst perfect-public equilibrium payoff can be attained by strategies that do not feature on-path group punishments, and the best and the worst can be jointly characterized as solutions to (3.48)-(3.51). Moreover, group punishments are sustainable and strictly improve welfare relative to a model where group punishments are not allowed.

In Figure 3.3.2, we provide a numerical illustration of how group punishments can increase the welfare of the team of oligopolists. In Figure 3.1, we fix the number of producers n to ten and plot the value of the best and worst perfect-public equilibria for each level of the discount factor δ. Note that in Figure 3.1, for any δ, values to the left of the static Nash equilibrium value (roughly 0.007) represent worst equilibrium values while values to the right represent best equilibrium values. The dashed line in Figure 3.1 shows these best and worst equilibrium values when group punishments are allowed, while the solid line shows these values when these punishments are not allowed. Since the dashed lines lie outside the solid lines, for all levels of the discount factor the model where taxes are allowed yields weakly higher best equilibrium payoffs than the model where taxes are not allowed. In particular, the repeated interaction between producers and the Principal leads to welfare gains for intermediate values of the discount factor, and no (or relatively small) gains when the discount factor is low or high.

For low values of δ, the Principal has weak incentives to levy the prescribed taxes. The continuation value that producers have to promise to the Principal for implementing such taxes is too to satisfy the feasibility constraint (3.50). As a result, very small or (approximately) no taxes can be sustained leading to small or (approximately) no welfare gains. On the other hand, for high values of δ the repeated interaction of producers is sufficient to guarantee the static most collusive level of output even in the absence of the Principal.

For intermediate levels of δ, the presence of the Principal increases welfare considerably. To illustrate the gains associated with sustainable group punishments (or taxes), Figure 3.2 illustrates the effect of the Principal's punishments on the level of output in the best equilibrium. Specifically, the solid line shows the percentage reduction in output in the best equilibrium which is obtained in our model relative to a model where group punishments are not allowed. Observe that our model features a most collusive output level as much as thirty percent lower than the model where group punishments are not allowed. To achieve these lower levels of output, which correspond to higher levels of welfare, the Principal reduces the value of the most profitable, static deviation by any of the producers by as much as 80%. This finding suggests that the role of the Principal in the oligopoly model is to decrease the common price to a level closer to the producer's marginal cost in case of a deviation, therefore reducing the value of deviations.

3.3.3 Substitutability and Price Externalities

In this section, we provide an overview of our additional results on how different degrees of interaction between oligpolistic producers can impact the effectiveness of group punishments. A full discussion of these results is provided in Appendix C.1.

The main point of departure of this section is the use of a new price function, which allows for different degrees of substitutability between producers' output. Specifically, we

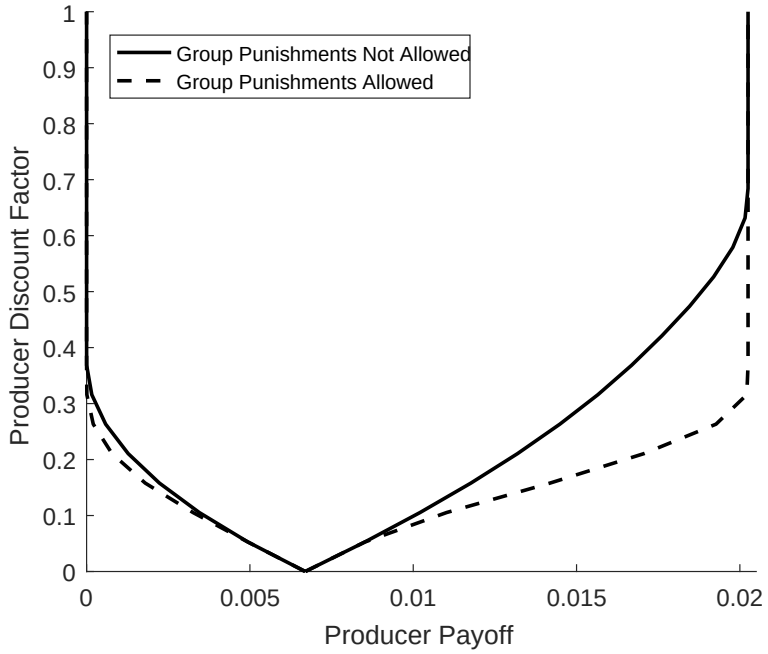


Figure 3.1: Equilibrium Value Sets

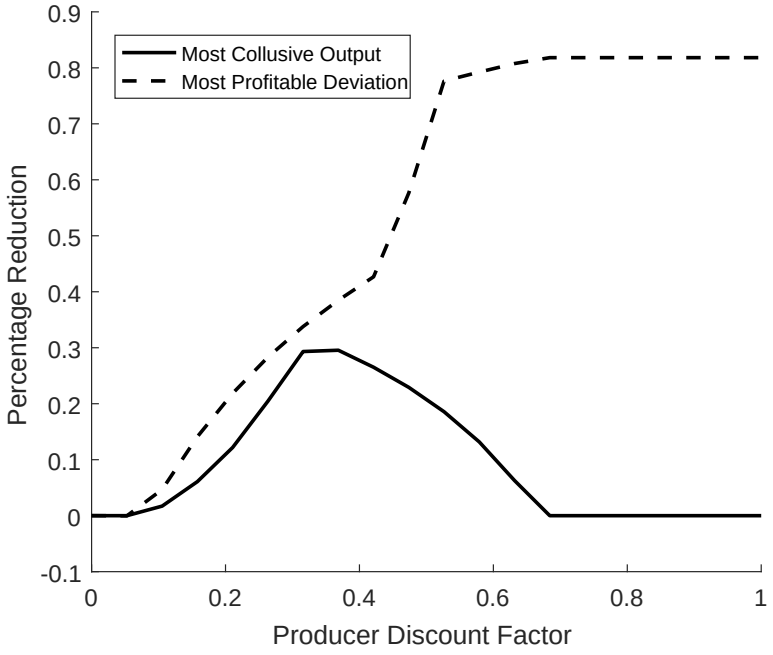


Figure 3.2: Impact of Group Punishments
Numerical illustration of the equilibrium value sets (panel (a)) and impact of group punishments on best equilibrium output and best deviation payoff from best equilibrium (panel (b)).

make the assumption that the inverse demand function for each producer i's output satisfies

$$p_i(q, \tau) = \alpha \frac{q_i^{\rho-1}}{\sum_{i=1}^{n} q_i^{\rho}} - \tau, \tag{3.52}$$

where $\alpha \in (0,1)$ and $\rho \in (0,1)$ are exogenous parameters, q_i is the quantity produced by producer i and τ is the tax chosen by the Principal. This price function arises naturally in an economy where consumers have Cobb-Douglas preferences over a bundle of individual producers' output and a numeraire good. In particular, the parameter α is a Cobb-Douglas parameter that governs the substitutability between the numeraire good and the bundle of producers' output, while the parameter ρ governs the degree of substitutability between each producer's output. Under this formulation, a higher level of ρ implies a higher degree of substitutability.

In the Appendix, we extend the analysis of the previous sections to the new inverse demand function (3.52), and we analyze the relationship between the usefulness of group punishments and the substitutability parameter ρ. Specifically, we ask how the effectiveness of taxes in improving welfare (relative to a model where taxes are not allowed) changes as the substitutability of producers' output changes. Our main result for this section shows that the effectiveness of taxes in improving welfare increases as the substitutability parameter ρ increases:

Proposition III.11. *Fix $\rho \in (0,1)$. For n sufficiently large, there exist a $\delta \in (0,1)$ and $\bar{\rho} > 0$ such that for all $\rho' \in (\rho, \bar{\rho})$, the welfare gains from allowing the Principal to implement group punishments are increasing in ρ'.*

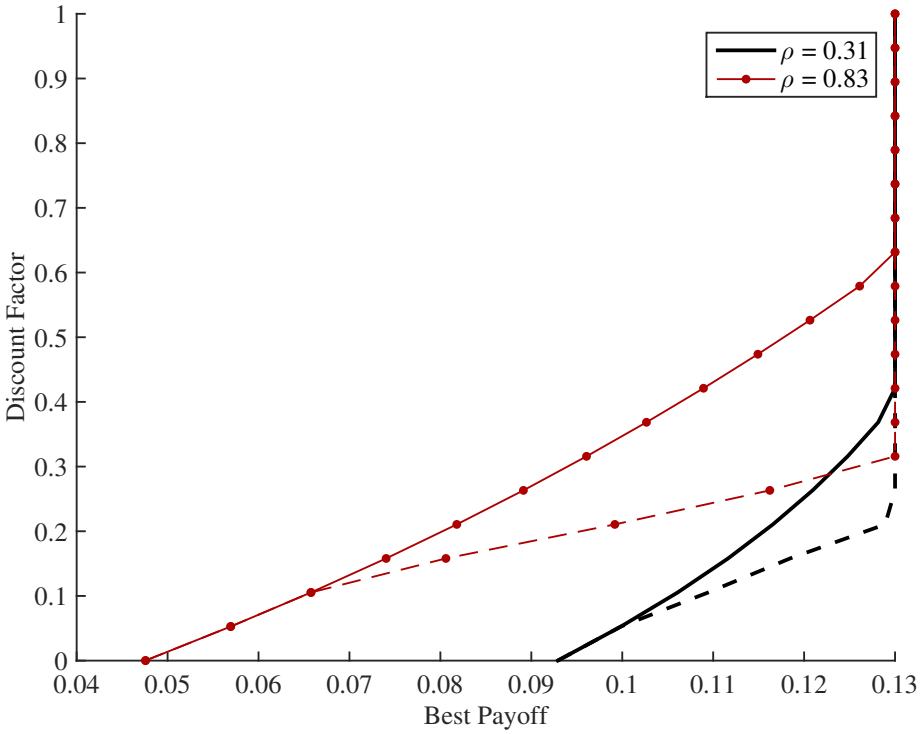
Proof. See Appendix C.1. □

The intuition behind the result of Proposition III.11 is that when goods become more substitutable, individual producers have higher incentives to deviate from their prescribed quantities because deviations have a lower negative impact on the common price. This increases the producers' incentives to over-produce and leads to lower equilibrium values,

but also increases the relative gains from group punishments relative to the model where these punishments are now allowed. In other words, when goods are more substitutable and deviations are more profitable, group punishments that deter these deviations increase welfare by more.

Finally, in Figure 3.3 we provide a numerical illustration of our result. The figure shows the value of the best equilibrium under a low value of the substitutability parameter ($\rho = 0.31$) and under a high value of the substitutability parameter ($\rho = 0.83$). As in Figure 3.1, the solid lines in Figure 3.3 represent the best equilibrium payoffs in the economies where group punishments are not allowed, and the dashed lines represent the equilibrium payoffs in the economies where group punishments are allowed. The difference between the dashed lines and the solid lines represent the welfare gains from allowing group punishments.

Figure 3.3: Value Sets with and without Group Punishments



Best equilibrium values for $\rho = 0.31$ and $\rho = 0.83$ when group punishments are not allowed (solid lines) and are allowed (dashed lines). In this example, we set $n = 5$, $\alpha = 0.7$ and $c = 0.1$.

The figure provides a clear illustration of our result that group punishments yields significantly larger increases in best equilibrium values when producers' output is more substitutable relative to when producers' output is less substitutable. For example, for a discount factor of roughly 0.4, with high degree of substitutability, the best equilibrium value when group punishments are not allowed is roughly 0.1 while it is roughly 0.13 when they are allowed, implying a 30% gain from group punishments. Instead, with a low degree of substitutability, the best equilibrium value when group punishment are not allowed is roughly 0.125 while it is roughly 0.13 when they are allowed, implying only a 4% gain from group punishments.

3.4 Conclusion

The potential for moral hazard is ubiquitous in team production settings and especially where the actions of individual team members are not perfectly observable. A widely accepted principle is that in these team production settings it is against the team's own interest to implement a group punishment when an individual deviation has occurred. An outsider is therefore needed to implement the team's first-best level of production.

In a generalized repeated team production model, we show that the team can always sustain self-imposed group punishments after aggregate outcomes are observed when team members' utility interacts in non-trivial ways with aggregate team outcomes. Moreover, we provide conditions under which the threat of these punishments improves the welfare of the team relative to a model where group punishments are not part of the team's action set. Using the repeated oligopoly model of *Abreu* (1986) as an application, we show that team self-imposed group punishments are most effective in improving team welfare when team members are sufficiently patient and when their contributions to the aggregate outcome are more substitutable.

Our theoretical results provide direct guidance for future applied and empirical research. In particular, our model predicts that team production environments featuring a strong inter-

action between aggregate outcomes and individual utilities are also environments where self-inflicted group punishments can provide large welfare gains to the team. Economic unions such as the European Union are particularly good examples of teams where team members have historically been tempted to deviate from their prescribed actions, and where aggregate team outcomes (e.g. common interest rates and exchange rates) interact in non-trivial ways with the individual utility of team members (e.g. individual output). Large corporations with multiple project managers are another setting to apply our model, especially since the presence of a non-benevolent top management lacking commitment to group punishments might exacerbate the moral hazard problem among individual project managers. Additional settings relevant to our analysis include environmental pacts, workplace management, and cartels. The analysis of the interaction between team members and the quantification of possible welfare gains from implementing group punishments in these settings constitutes in our opinion areas of fruitful future research.

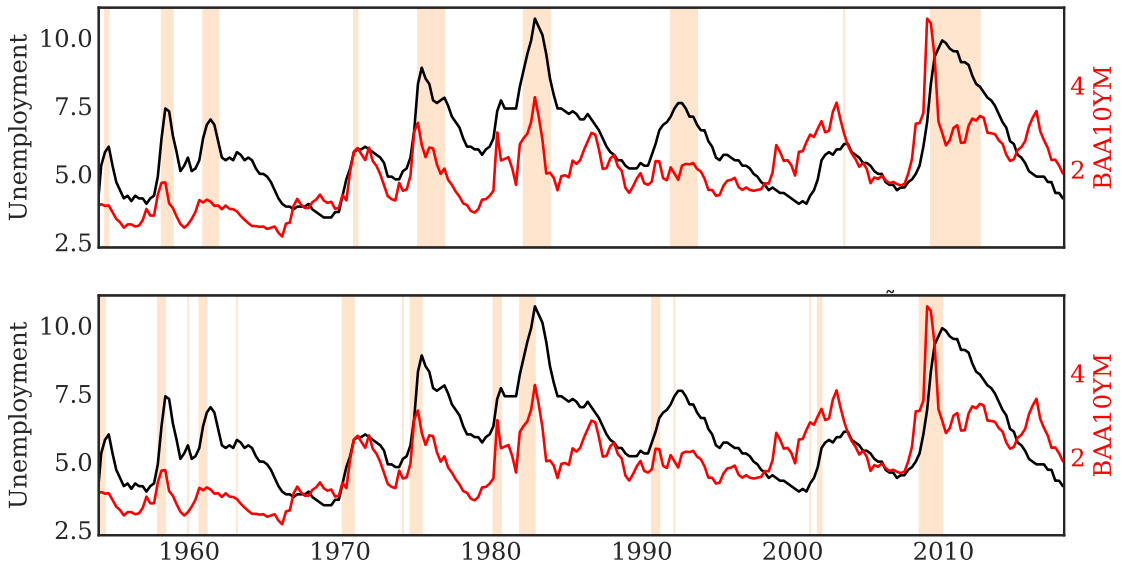
APPENDICES

APPENDIX A

Appendix for Credit Market Search

A.1 Identifying Recessions

Figure A.1: Discrete Economic States as Determined by $\tilde{\mathcal{U}}_t$



A.2 Representative Household

The representative household's problem is:

$$
\begin{aligned}
H_t(\mathcal{N}_t, A_t) &= \max_{C_t, A_t} \left[u(C_t) + l\mathcal{U}_t\right] + \beta \mathbb{E}_t \left[H_{t+1}(\mathcal{N}_{t+1} A_{t+1})\right] && \text{(A.1)}\\
\text{subject to} \quad & W_t \mathcal{N}_t + b\mathcal{U}_t + A_{t-1}(1 + r_{t-1}) + D_t^S + D_t^B = C_t + T_t + A_t && \text{(A.2)}\\
& \mathcal{N}_{t+1} = \left(1 - s^C\right)\left(1 - s^L\right)\mathcal{N}_t + q(\theta_t)\mathcal{V}_t && \text{(A.3)}\\
& \mathcal{U}_{t+1} = (1 - f(\theta_t))\,\mathcal{U}_t + \left(s^C + (1 - s^C)s^L\right)\mathcal{N}_t && \text{(A.4)}
\end{aligned}
$$

Let λ_t denote the Lagrange multiplier for the household budget constraint. The first order conditions w.r.t. to C_t is:

$$
u_c(C_t) = \lambda_t
$$

The first order condition w.r.t. to A_t is:

$$
\begin{aligned}
0 &= -\lambda_t + \beta \mathbb{E}_t \left[\frac{\partial H_{t+1}}{\partial A_t}\right] && \text{(A.5)}\\
1 &= \beta \mathbb{E}_t \left[\frac{u_c(C_{t+1})}{u_c(C_t)}\,(1 + r_t)\right] && \text{(A.6)}
\end{aligned}
$$

A.2.1 Marginal values of employed and unemployed household members

The total derivative of the household's value function w.r.t $\mathcal{N}_t$ is given by:

$$
\begin{aligned}
H_{N_t} &= W_t \lambda_t + \beta \mathbb{E}_t \left(\frac{\partial H_{t+1}}{\partial \mathcal{N}_{t+1}}\frac{\partial \mathcal{N}_{t+1}}{\partial \mathcal{N}_t} + \frac{\partial H_{t+1}}{\partial \mathcal{U}_{t+1}}\frac{\partial \mathcal{U}_{t+1}}{\partial \mathcal{N}_t}\right)\\
H_{N_t} &= W_t \lambda_t + \beta \mathbb{E}_t \left[(1 - s^C)(1 - s^L) H_{N_{t+1}} + \left(s^C + (1 - s^C)s^L\right) H_{U_{t+1}}\right]\\
\frac{H_{N_t}}{\lambda_t} &= W_t + \beta \mathbb{E}_t \frac{1}{\lambda_t} \left[(1 - s^C)(1 - s^L) H_{N_{t+1}} + \left(s^C + (1 - s^C)s^L\right) H_{U_{t+1}}\right]\\
\frac{H_{N_t}}{\lambda_t} &= W_t + \beta \mathbb{E}_t \left(\frac{\lambda_{t+1}}{\lambda_t}\left[(1 - s^C)(1 - s^L)\frac{H_{N_{t+1}}}{\lambda_{t+1}} + \left(s^C + (1 - s^C)s^L\right)\frac{H_{U_{t+1}}}{\lambda_{t+1}}\right]\right)
\end{aligned}
$$

The total derivative of the household's value function w.r.t $\mathcal{U}_t$ is given by:

$$\begin{aligned}
H_{U_t} &= b\lambda_t + l + \beta\mathbb{E}_t\left(\frac{\partial H_{t+1}}{\partial \mathcal{N}_{t+1}}\frac{\partial \mathcal{N}_{t+1}}{\partial \mathcal{U}_t} + \frac{\partial H_{t+1}}{\partial \mathcal{U}_{t+1}}\frac{\partial \mathcal{U}_{t+1}}{\partial \mathcal{U}_t}\right) \\
H_{U_t} &= b\lambda_t + l + \beta\mathbb{E}_t\left[f(\theta_t)H_{N_{t+1}} + (1-f(\theta_t))H_{U_{t+1}}\right] \\
H_{U_t} &= b\lambda_t + l + \beta\mathbb{E}_t\left[f(\theta_t)H_{N_{t+1}} + (1-f(\theta_t))H_{U_{t+1}}\right] \\
\frac{H_{U_t}}{\lambda_t} &= b + \frac{l}{\lambda_t} + \beta\mathbb{E}_t\left(\frac{\lambda_{t+1}}{\lambda_t}\left[f(\theta_t)\frac{H_{N_{t+1}}}{\lambda_{t+1}} + (1-f(\theta_t))\frac{H_{U_{t+1}}}{\lambda_{t+1}}\right]\right)
\end{aligned}$$

A.3 Repayment to Creditors

The expected repayment rule that solves the Nash bargaining must satisfy the sharing rule:

$$(1-\alpha_C)B_{l,t} = \alpha_C S_{l,t} \tag{B.1}$$

Expanding both sides of the equality using equations (1.12) and (1.6), we have:

$$\begin{aligned}
(1-\alpha_C)\left[-\gamma + \left(1-s^C\right)\mathbb{E}_t M_{t+1}\left[q_t B_{g,t+1}\right]\right] &= \alpha_C\left(1-s^C\right)\mathbb{E}_t M_{t+1}\left[q_t S_{g,t+1}\right] \\
&\quad + \left(1-s^C\right)(1-q_t)\mathbb{E}_t M_{t+1}\left[\alpha_C S_{l,t+1}\right] \\
&\quad - \left(1-s^C\right)(1-q_t)\mathbb{E}_t M_{t+1}\left[(1-\alpha_C)B_{l,t+1}\right]
\end{aligned}$$

$$(1-\alpha_C)\left[-\gamma + \left(1-s^C\right)\mathbb{E}_t M_{t+1}\left[q_t B_{g,t+1}\right]\right] = \alpha_C\left(1-s^C\right)\mathbb{E}_t M_{t+1}\left[q_t S_{g,t+1}\right]$$

$$(1-\alpha_C)\,\mathbb{E}_t M_{t+1}\left[B_{g,t+1}\right] = (1-\alpha_C)\frac{\gamma}{q_t\left(1-s^C\right)} + \alpha_C\mathbb{E}_t M_{t+1}\left[S_{g,t+1}\right]$$

$$\begin{aligned}
&(1-\alpha_C)\,\mathbb{E}_t M_{t+1}\left[\Psi_{t+1}\right. \\
&\left. + \left(1-s^C\right)M_{t+2}\left[(1-s^L)B_{g,t+2} + s^L B_{l,t+2}\right]\right] &= (1-\alpha_C)\frac{\gamma}{q_t\left(1-s^C\right)} \\
&\quad + \alpha_C\mathbb{E}_t M_{t+1}\left[X_{t+1} - W_{t+1} - \Psi_{t+1}\right. \\
&\quad \left. + \left(1-s^C\right)M_{t+2}\left[\left(1-s^L\right)S_{g,t+2} + s^L S_{l,t+2}\right]\right]
\end{aligned}$$

$$
\begin{aligned}
\mathbb{E}_t M_{t+1}\left[\Psi_{t+1}\right] &= (1-\alpha_C)\frac{\gamma}{q_t\left(1-s^C\right)} + \alpha_C \mathbb{E}_t M_{t+1}\left[X_{t+1}-W_{t+1}\right] \\
&\quad +\alpha_C \mathbb{E}_t M_{t+1}\left[\left(1-s^C\right) M_{t+2}\left[\left(1-s^L\right) S_{g,t+2} + s^L S_{l,t+2}\right]\right] \\
&\quad -(1-\alpha_C)\,\mathbb{E}_t M_{t+1}\left(1-s^C\right) M_{t+2}\left[(1-s^L)B_{g,t+2} + s^L B_{l,t+2}\right] \\
\mathbb{E}_t M_{t+1}\left[\Psi_{t+1}\right] &= (1-\alpha_C)\frac{\gamma}{q_t\left(1-s^C\right)} + \alpha_C \mathbb{E}_t M_{t+1}\left[X_{t+1}-W_{t+1}\right] \\
&\quad +\left(1-s^C\right)\mathbb{E}_t M_{t+1}\left[M_{t+2}\left[\left(1-s^L\right)\left(\alpha_C S_{g,t+2} - (1-\alpha_C)\,B_{g,t+2}\right)\right.\right. \\
&\quad \left.\left. +s^L\left(\alpha_C S_{l,t+2} - (1-\alpha_C)\,B_{lt+2}\right)\right]\right] \\
\mathbb{E}_t M_{t+1}\left[\Psi_{t+1}\right] &= \alpha_C \mathbb{E}_t M_{t+1}\left[X_{t+1}-W_{t+1}\right] + (1-\alpha_C)\frac{\gamma}{q_t\left(1-s^C\right)} \\
&\quad +\left(1-s^C\right)\mathbb{E}_t M_{t+1}\left[\left(1-s^L\right) M_{t+2}\left[\left(\alpha_C S_{g,t+2} - (1-\alpha_C)\,B_{g,t+2}\right)\right]\right] \\
\mathbb{E}_t M_{t+1}\left[\Psi_{t+1}\right] &= \alpha_C \mathbb{E}_t M_{t+1}\left[X_{t+1}-W_{t+1}\right] + (1-\alpha_C)\frac{\gamma}{q_t\left(1-s^C\right)} \\
&\quad +\left(1-s^L\right)\mathbb{E}_t M_{t+1}\left[\left(1-s^C\right) M_{t+2}\left[\left(\alpha_C S_{g,t+2} - (1-\alpha_C)\,B_{g,t+2}\right)\right]\right]
\end{aligned}
$$

Expanding the Nash sharing rule leads to:

$$
\begin{aligned}
0 &= \alpha_C S_{l,t} - (1-\alpha_C) B_{l,t} \\
0 &= \alpha_C\left(1-s^C\right)\mathbb{E}_t M_{t+1}\left[q_t S_{g,t+1} + (1-q_t) S_{l,t+1}\right] \\
&\quad -(1-\alpha_C)\left[-\gamma + \left(1-s^C\right)\mathbb{E}_t M_{t+1}\left[q_t B_{g,t+1} + (1-q_t) B_{l,t+1}\right]\right] \\
0 &= (1-\alpha_C)\,\gamma + q_t\left(1-s^C\right)\mathbb{E}_t M_{t+1}\left[\alpha_C S_{g,t+1} - (1-\alpha_C)\,B_{gt+1}\right] \\
&\quad +\left(1-s^C\right)(1-q_t)\mathbb{E}_t M_{t+1}\left[\alpha_C S_{l,t+1} - (1-\alpha_C) B_{l,t+1}\right] \\
-(1-\alpha_C)\frac{\gamma}{q_t} &= \left(1-s^C\right)\mathbb{E}_t M_{t+1}\left[\alpha_C S_{g,t+1} - (1-\alpha_C)\,B_{gt+1}\right]
\end{aligned}
$$

which, iterated one period forward, can be substituted into the previous expression to yield:

$$
\mathbb{E}_t M_{t+1}\left[\Psi_{t+1}\right] = \alpha_C \mathbb{E}_t M_{t+1}\left[X_{t+1}-W_{t+1}\right] + (1-\alpha_C)\left[\frac{\gamma}{q_t\left(1-s^C\right)} - \left(1-s^L\right)\mathbb{E}_t\left[M_{t+1}\frac{\gamma}{q_{t+1}}\right]\right] \tag{B.2}
$$

A.4 Job creation condition

$$\begin{aligned}
F_{lt} &= -\gamma + \left(1 - s^C\right) \mathbb{E}_t M_{t+1} \left[q_t F_{g,t+1} + (1 - q_t) F_{l,t+1}\right] \\
K(\phi_t) &= -\gamma + \left(1 - s^C\right) \mathbb{E}_t M_{t+1} \left[q_t F_{g,t+1} + (1 - q_t) K(\phi_{t+1})\right] \\
K(\phi_t) - \left(1 - s^C\right)(1 - q_t)\, \mathbb{E}_t M_{t+1} K(\phi_{t+1}) &= -\gamma + q_t \left(1 - s^C\right) \mathbb{E}_t \left[M_{t+1} F_{g,t+1}\right] \\
\frac{K(\phi_t) - \left(1 - s^C\right)(1 - q_t)\, \mathbb{E}_t M_{t+1} K(\phi_{t+1}) + \gamma}{q_t \left(1 - s^C\right)} &= \mathbb{E}_t \left[M_{t+1} F_{g,t+1}\right] \\
\frac{\Gamma_t}{q_t} &= \mathbb{E}_t \left[M_{t+1} F_{g,t+1}\right]
\end{aligned}$$

where $\Gamma_t = \frac{K(\phi_t) - \left(1 - s^C\right)(1 - q_t)\mathbb{E}_t M_{t+1} K(\phi_{t+1}) + \gamma}{(1 - s^C)}$, and then

$$\begin{aligned}
F_{gt} &= X_t - W_t + \left(1 - s^C\right) \mathbb{E}_t M_{t+1} \left[\left(1 - s^L\right) F_{g,t+1} + s^L K(\phi_{t+1})\right] \\
F_{gt} &= X_t - W_t + \left(1 - s^C\right) \left[\left(1 - s^L\right) \frac{\Gamma_t}{q_t} + s^L \mathbb{E}_t M_{t+1} K(\phi_{t+1})\right]
\end{aligned}$$

which yield the job creation condition:

$$\frac{\Gamma_t}{q_t} = \mathbb{E}_t M_{t+1} \left[X_{t+1} - W_{t+1} + \left(1 - s^C\right) \left[\left(1 - s^L\right) \frac{\Gamma_{t+1}}{q_{t+1}} + s^L K(\phi_{t+1})\right]\right]$$

A.5 Nash Bargained Wage

The wage is bargained between an individual worker and the creditor-project pair, or with a marginal job for the firm. It solves the following problem:

$$argmax \left(\frac{H_{Nt} - H_{Ut}}{\lambda_t}\right)^{\alpha_L} (F_{gt} - F_{lt})^{1 - \alpha_L}$$

and thus must satisfy the labor match surplus sharing rule:

$$(1 - \alpha_L) \left(\frac{H_{Nt} - H_{Ut}}{\lambda_t}\right) = \alpha_L \left(F_{gt} - F_{lt}\right)$$

Begin with the right hand side:

$$\begin{aligned}\left(\frac{H_{Nt}-H_{Ut}}{\lambda_t}\right) &= W_t+\mathbb{E}_t\left(M_{t+1}\left[(1-s^C)(1-s^L)\frac{H_{N_{t+1}}}{\lambda_{t+1}}+\left(s^C+(1-s^C)s^L\right)\frac{H_{U_{t+1}}}{\lambda_{t+1}}\right]\right)\\ &\quad -Z_t+-\mathbb{E}_t\left(M_{t+1}\left[f(\theta_t)\frac{H_{N_{t+1}}}{\lambda_{t+1}}+(1-f(\theta_t))\frac{H_{U_{t+1}}}{\lambda_{t+1}}\right]\right)\\ &= W_t-Z_t-f(\theta_t)\mathbb{E}_t\left(M_{t+1}\left(\frac{H_{N_{t+1}}-H_{U_{t+1}}}{\lambda_{t+1}}\right)\right)\\ &\quad +(1-s^C)(1-s^L)\mathbb{E}_t\left(M_{t+1}\left[\frac{H_{N_{t+1}}-H_{U_{t+1}}}{\lambda_{t+1}}\right]\right)\end{aligned}$$

and working with the left hand side;

$$\begin{aligned}(F_{gt}-K(\phi^*)) &= \left(X_t-W_t+\left(1-s^C\right)\mathbb{E}_tM_{t+1}\left[\left(1-s^L\right)F_{g,t+1}+s^LK(\phi^*)\right]-K(\phi^*)\right)\\ &= \left(X_t-W_t+\left(1-s^C\right)\mathbb{E}_tM_{t+1}\left[\left(1-s^L\right)(F_{g,t+1}-K(\phi*))+K(\phi^*)\right]-K(\phi^*)\right)\end{aligned}$$

Substituting into the sharing rule, we have:

$$(1-\alpha_L)\left(W_t-Z_t-f(\theta_t)\mathbb{E}_t\left(M_{t+1}\left(\frac{H_{N_{t+1}}-H_{U_{t+1}}}{\lambda_{t+1}}\right)\right)\right) = \alpha_L\left(X_t-W_t+\left[\left(1-s^C\right)\mathbb{E}_tM_{t+1}-1\right]K(\phi^*)\right)$$

and

$$\begin{aligned}W_t &= \alpha_L\left(X_t+\left[\left(1-s^C\right)\mathbb{E}_tM_{t+1}-1\right]K(\phi^*)\right)+(1-\alpha_L)Z_t\\ &\quad +(1-\alpha_L)f(\theta_t)\mathbb{E}_t\left(M_{t+1}\left(\frac{H_{N_{t+1}}-H_{U_{t+1}}}{\lambda_{t+1}}\right)\right)\\ &\quad =\alpha_L\left(X_t+\left[\left(1-s^C\right)\mathbb{E}_tM_{t+1}-1\right]K(\phi^*)\right)+(1-\alpha_L)Z_t\\ &\quad +\alpha_Lf(\theta_t)\mathbb{E}_t\left[M_{t+1}\left(F_{gt+1}-K(\phi^*)\right)\right]\end{aligned}$$

Use

$$\frac{K(\phi^*)\left[1-\left(1-s^C\right)(1-q_t)\mathbb{E}_tM_{t+1}\right]+\gamma}{q_t\left(1-s^C\right)} = \mathbb{E}_t\left[M_{t+1}F_{g,t+1}\right]$$

to obtain

$$
\begin{aligned}
W_t &= \alpha_L \left(X_t + \left[\left(1-s^C\right)\mathbb{E}_t M_{t+1} - 1\right] K(\phi^*)\right) + (1-\alpha_L) Z_t \\
t &\quad +\alpha_L f(\theta_t)\mathbb{E}_t \left[\frac{K(\phi^*)\left[1-\left(1-s^C\right)(1-q_t)\mathbb{E}_t M_{t+1}\right]+\gamma}{q_t\left(1-s^C\right)} - M_{t+1}K(\phi^*)\right] \\
&= \alpha_L \left(X_t + \left[\left(1-s^C\right)\mathbb{E}_t M_{t+1} - 1\right] K(\phi^*)\right) + (1-\alpha_L) Z_t \\
&\quad +\alpha_L \theta_t \mathbb{E}_t \left[\frac{K(\phi^*)\left[1-\left(1-s^C\right)(1-q_t)\mathbb{E}_t M_{t+1}\right]+\gamma}{\left(1-s^C\right)} - q_t M_{t+1}K(\phi^*)\right] \\
&= \alpha_L X_t + (1-\alpha_L) Z_t \\
&\quad +\alpha_L \left[\left[\left(1-s^C\right)\mathbb{E}_t M_{t+1} - 1\right] K(\phi^*) + \right. \\
&\quad \left. \theta_t \left[\frac{K(\phi^*)\left[1-\left(1-s^C\right)(1-q_t)\mathbb{E}_t M_{t+1}\right]+\gamma}{\left(1-s^C\right)} - q_t M_{t+1}K(\phi^*)\right]\right] \\
&= \alpha_L X_t + (1-\alpha_L) Z_t \\
&\quad +\alpha_L \left[\left[\left(1-s^C\right)\mathbb{E}_t M_{t+1} - 1\right] K(\phi^*) \right. \\
&\quad \left. +\theta_t \left[\frac{\gamma}{\left(1-s^C\right)} + \left[\frac{1}{\left(1-s^C\right)} - (1-q_t)\mathbb{E}_t M_{t+1}\right] K(\phi^*) - q_t M_{t+1}K(\phi^*)\right]\right] \\
W_t &= \alpha_L \left(X_t + \theta_t \left[\frac{\gamma}{\left(1-s^C\right)} + \left[\frac{r_t + s^C}{\left(1-s^C\right)(1+r_t)}\right] K(\phi^*)\right]\right) + (1-\alpha_L) Z_t - \alpha_L \left[\frac{r_t+s^C}{1+r_t}\right] K(\phi^*)
\end{aligned}
$$

APPENDIX B

Appendix for Market Size and Market Concentration

B.1 Appendix: Numerical Solution Methods

In this section I describe my solution method in detail. Code used to solve the model, along with documentation, is included at `https://github.com/btengels.com`. My solution algorithm follows the steps below:

1. **Establish a discrete state space and transition probability matrix for z_t.** I create evenly-spaced grids of values for k_t, Q_t, z_t containing 20, 17, and 13 points respectively. The ranges of these grids are large enough such that further expansion has no noticeable impact on the model's results. The transition probability matrix for z_t is computed using the method of *Rouwenhorst* (1995) which is ideal for simulating autoregressive processes with high levels of persistence.

2. **Establish a discrete action space for choice variable i_t.** I create an evenly-spaced grid of potential i_t values. The lower limit is 0 and the upper limit is chosen such that increasing the limit further has no impact on the model's results.

3. **Create an initial guess for value function J.** I use a matrix zeros with three dimensions corresponding to the sizes of the grids for k, Q and z.

4. **For every state (K, Q, z) and for every choice i:**

    - Compute the contemporaneous payoff $\pi(k_t, Q_t, z_t, i_t)$ using equation (2.6).
    - Compute k_{t+1} given values for K and i and equation (2.4).

- Use the value of k_{t+1} to interpolate the value function J for potential future states $(k_{t+1}, Q_{t+1}, z_{t+1})$
- Use the transition probability matrix and the expectation rule in equation **??** to compute the expected value of J_{t+1} given the current state and choice of i
- Determine the optimal choice of i based on the sum of current payoffs $\pi(k_t, Q_t, z_t, i_t)$ and continuation values $E_t\left[J(k_{t+1}, Q_{t+1}, z_{t+1})\right]$
- Determine whether choosing $i_t > 0$ is better or worse than exiting and choosing $i_t = 0$. This occurs when $E_t\left[J_{exit}(k_{t+1}, Q_{t+1}, z_{t+1})\right] > E_t\left[J(k_{t+1}, Q_{t+1}, z_{t+1})\right])$
- Save the optimal decision for i_t, e_t at each state

5. **Update the value function.** Use the values of J resulting from the optimal choice of i at each state to update the value function $J(k, Q, z)$

6. **Iterate until convergence.** Return to Step 4 and repeat until the value function J converges.

B.2 Appendix: Details on Weighting Functions

Prices in my model are set via an inverse demand function that considers the output from competitors in typical Cournot-oligopoly fashion. I add a notion of distance between competitors and the affect of their output on competitors' prices through a weighting function $w(\cdot)$. This weighting function is decreasing with the distance between two firms i and j, such that nearby firms have a larger affect on one another. The only requirements I assume for $w(\cdot)$ are $\frac{\partial w(x_i,x_j;h)}{\partial d(x_i,x_j)} \leq 0$, $w(\cdot) \in [0,1]$, and $w(x_i, x_j) = w(x_j, x_i)$. All of these are satisfied by the family of symmetric kernel functions.

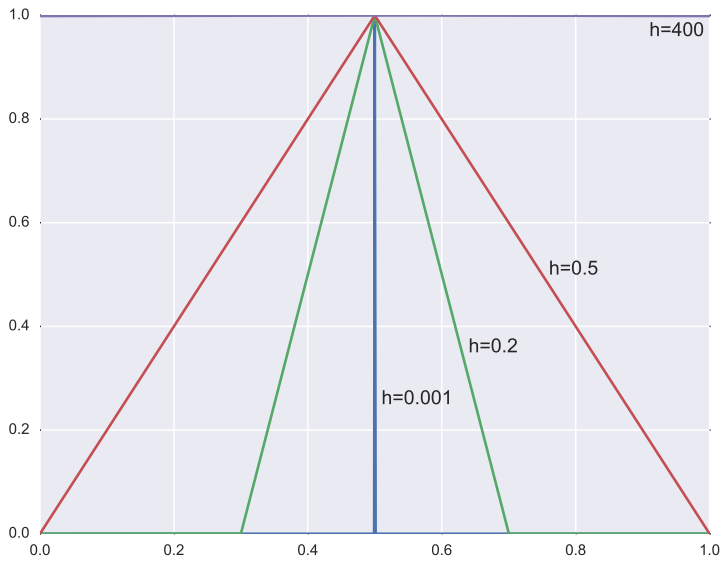
The main kernel function used throughout the paper is the "triangular" kernel, given by

$$w(x_i, x_j; h) = \max\left\{0, \left[1 - \frac{d(x_i, x_j)}{h}\right]\right\}. \tag{D.1}$$

The parameter h, called the "bandwidth" of the function, determines the range of distances which will receive positive weights. With the triangular kernel, a firm located at x_i will be affected by firms in the interval $(x - h, x + h)$. Figure B.1 plots $w(x_i, x_j; h)$ for $x_i = .5$ and $x_j \in [0, 1]$ for different values of h.

When $h \to 0$, the interval $(x_i - h, x_j + h)$ approaches a single point and $w(x_i, x_j; h) = 0$ for all $x_i \neq x_j$. Conversely, when $h \to \infty$, the term $\frac{d(x_i,x_j)}{h} \to 0$ and $w(x_i, x_j; h) \to 1$ for any pair of firms x_i, x_j. In trying different functions, including parabolic and Gaussian

Figure B.1: Triangular Weighting Function for Different h Values



kernels, the shape of the function does not seem to matter as much as the value of h. I use a triangular kernel throughout my analysis due to its simplicity and interpretability.

B.3 Appendix: Details on Inverse Demand Function

In Section **??**, I introduce a pricing function that adjusts based on the number of active firms in the economy. The goal is to prevent the aggregate economy from shrinking as firms exit. In a hypothetical scenario where prices are fixed at some price $\bar{p}$ and half of active firms exit, the surviving firms would clear the market at $\bar{p}$ by producing twice as much output as before.

$$\bar{p} = \chi - \left(\frac{N_t}{N_1}\right) Q \tag{D.2}$$

$$\left(\frac{N_t}{N_1}\right) Q = (\chi - \bar{p}) \tag{D.3}$$

$$Q = (\chi - \bar{p}) \left(\frac{N_1}{N_t}\right) \tag{D.4}$$

Let $\bar{Q}$ denote the value of Q at period 1 when $N_t = N_1$. If at some future period $N_t = .5N_1$, then equation (D.4) becomes $Q = (\chi - \bar{p})\, 2 = 2\bar{Q}$ and the amount of output by the firm required to clear the market at $\bar{p}$ has doubled. The same is for any ratio between

N_t and N_1, such that the aggregate market size will increase proportionally to the share of firms exiting each period.

B.4 Appendix: Change in CR4 for Select 4-Digit NAICS Codes

NAICS	Description	CR4 2002	CR4 2012	Rev. Growth
5172	Wireless telecommunications carriers (except s...	61.7	89.1	328.4
4512	Book, periodical, and music stores	48.3	66.1	76.0
4422	Home furnishings stores	20.9	36.4	168.5
4461	Health and personal care stores	45.7	60.0	199.8
5174	Satellite telecommunications	34.6	48.1	164.5
4441	Building material and supplies dealers	42.0	53.8	142.9
4541	Electronic shopping and mail-order houses	18.7	30.2	433.3
5191	Other information services	30.7	41.9	2619.6
4431	Electronics and appliance stores	44.3	54.1	152.2
4539	Other miscellaneous store retailers	13.2	22.9	210.1
4421	Furniture stores	8.1	17.3	207.5
4413	Automotive parts, accessories, and tire stores	21.2	30.3	192.0
4442	Lawn and garden equipment and supplies stores	10.4	18.4	223.2
4542	Vending machine operators	20.7	28.2	113.4
4521	Department stores	66.4	73.2	88.6
4471	Gasoline stations	8.2	13.3	360.6
4529	Other general merchandise stores	78.8	82.7	216.4
4511	Sporting goods, hobby, and musical instrument ...	24.2	27.4	146.5
4532	Office supplies, stationery, and gift stores	45.9	48.8	91.6
4533	Used merchandise stores	9.9	12.7	214.4
4483	Jewelry, luggage, and leather goods stores	22.3	24.9	137.7
4412	Other motor vehicle dealers	4.2	6.4	154.0
5112	Software publishers	39.5	41.4	173.0
4453	Beer, wine, and liquor stores	8.3	10.1	184.3
5179	Other telecommunications	31.4	32.6	1825.4
5151	Radio and television broadcasting	39.1	39.8	132.5
4411	Automobile dealers	5.3	5.9	119.3
5111	Newspaper, periodical, book, and directory pub...	13.6	14.0	75.0
4543	Direct selling establishments	11.0	11.2	129.4
4531	Florists	1.7	1.6	66.4
4452	Specialty food stores	6.8	5.9	117.6
4481	Clothing stores	28.0	27.0	138.2
4451	Grocery stores	31.0	29.8	129.7
5121	Motion picture and video industries	37.4	34.9	119.5
5221	Depository credit intermediation	24.2	20.8	68.2
5152	Cable and other subscription programming	63.9	58.9	230.7
4482	Shoe stores	39.9	34.3	114.9
5171	Wired telecommunications carriers	59.7	51.3	103.4
5122	Sound recording industries	60.9	51.0	61.2
5182	Data processing, hosting, and related services	33.7	15.9	95.7

APPENDIX C

Appendix for Group Punishments

C.1 Substitutability and Price Externalities

In this Appendix extend the analysis of Section 3.3 by allowing for a rich degree of price externalities. In particular, we analyze the degree to which the ability of the Principal to impose group punishments improves social welfare when varying the degree of substitutability between the output of individual producers. In Section C.1.1, we introduce a new pricing function which admits a variable degree of substitution across producers' goods and derive equilibrium outcomes of the stage game. In Section C.1.2, we develop a recursive formulation of the infinite-horizon game and show that the usefulness of group punishments increases as goods become more substitutable.

C.1.1 Stage Game

We generalize the price function by assuming that consumers have Cobb-Douglas preferences over a bundle of individual producers' output and a numeraire good, and that these consumers face taxes τ on purchases of each producers' output. In this economy, the inverse demand function for each producer i's output satisfies

$$p_i(q, \tau) = \alpha \frac{q_i^{\rho-1}}{\sum_{i=1}^{n} q_i^{\rho}} - \tau, \tag{D.1}$$

where $\alpha \in (0,1)$ and $\rho \in (0,1)$. Here, the parameter α is a Cobb-Douglas parameter that governs the substitutability between the numeraire good and the bundle of producers' output, while the parameter ρ governs the degree of substitutability between each producer's output. Under this formulation, a higher level of ρ implies a higher degree of substitutability.

With prices specified in (D.1), each producer i obtains a static payoff given by

$$u_i(q,\tau) = \alpha \frac{q_i^\rho}{\sum_{i=1}^n q_i^\rho} - \tau q_i - c q_i, \tag{D.2}$$

while the Principal obtains a static payoff given by

$$w(Q,\tau) = \sum_{i=1}^n p_i(q,\tau) q_i \tag{D.3}$$

$$= \alpha - \tau Q. \tag{D.4}$$

As in the case of a linear inverse demand function, after observing any level Q, the Principal optimally chooses $\tau(Q) = 0$ in the stage game. We impose a restriction on the strategy set of each producer which requires strictly positive production. Formally, we restrict $q_i \in \left[\underline{q}, \infty\right]$ with $\underline{q} < q_i^N$. Under this restriction, the level of output that maximizes joint profits in the stage game satisfies

$$q_i^m = \arg\max_{q_i \geq \underline{q}} \left(\frac{\alpha}{n} - c q_i \right) \tag{D.5}$$

$$= \underline{q}. \tag{D.6}$$

Next, to solve for the unique perfect-public equilibrium of the stage game q_i^N, we note that for each q_{-i}, producer i solves

$$\max_{q_i} \alpha \frac{q_i^\rho}{q_i^\rho + \sum_{-i} q_{-i}^\rho} - c q_i.$$

It is straightforward to show that the unique perfect-public equilibrium of the stage game is

$$q_i^N = \frac{n-1}{n^2} \frac{\alpha \rho}{c}. \tag{D.7}$$

C.1.2 Infinitely-Repeated Game

We focus on characterizing *strongly* symmetric perfect-public equilibria. We denote by $u(q,\tau)$ the producer's payoff and by $w(Q,\tau)$ the Principal's payoff, and after appealing to the one-shot deviation principle, we proceed to characterize the best and worst perfect-public equilibria of the repeated game. Under the inverse demand function (D.1), for a given level of the worst equilibrium payoff $\underline{v}$ the best equilibrium payoff $\bar{v}$ solves

$$\bar{v} = \max_q u(q,0),$$

subject to, for all q',

$$u(q,0) \geq (1-\delta)\, g(q', q, \tau(q' + (n-1)\, q)) + \delta \underline{v}, \tag{D.8}$$

$$\bar{v} \geq \frac{1-\delta}{\delta} \frac{1}{n} \left[w(q' + (n-1)\, q, 0) - w(q' + (n-1)\, q, \tau(q' + (n-1)\, q)) \right] + \underline{v}, \tag{D.9}$$

where $g(q', q, \tau(q' + (n-1)\, q))$ now satisfies

$$\begin{aligned} g(q', q, \tau(q' + (n-1)\, q)) = {} & u(q', q, \tau(q' + (n-1)\, q)] + \frac{1}{n} w(q' + (n-1)\, q, 0) \\ & - \frac{1}{n} w(q' + (n-1)\, q, \tau(q' + (n-1)\, q)). \end{aligned} \tag{D.10}$$

As in the previous section, we define the maximum payoff that can be achieved by a producer by deviating to q' when the others are producing q as $\hat{g}(q, \tau(\cdot))$. This maximum payoff satisfies

$$\hat{g}(q, \tau(\cdot)) = \max_{q'} g(q', q, \tau(q' + (n-1)q)).$$

In the next lemma, we show that as long as the prescribed output is larger than the static Nash equilibrium output, the maximum deviation payoff $\hat{g}(q, \tau(\cdot))$ is minimized when the Principal levies no taxes (i.e., when $\tau = 0$).

Lemma C.1. *$\hat{g}(q, \tau(\cdot)) \geq \hat{g}(q, \tau = 0)$ when $q \geq q^N$.*

Proof. See Appendix C.2.2.1. □

Given Lemma C.1, the key propositions of Section 3.3 immediately extend to the environment with imperfectly substitutable goods. Here, we explore how the usefulness of group punishments in improving welfare depends on the degree of substitutability between individual producers' output. We start by showing in the following lemma that when the number of producers n is sufficiently large, the best equilibrium level of output of the model where taxes are not allowed is increasing in the substitutability parameter ρ.

Lemma C.2. *For n sufficiently large, $d\bar{q}^A/d\rho > 0$.*

Proof. See Appendix C.2.2.2. □

The intuition behind this lemma is that when output is more substitutable the negative impact of an individual producer's output on the common price is lower. This increases producers' incentives to over-produce, and leads to higher levels of production and lower equilibrium values in the best equilibrium.

Finally, in the following proposition we formalize our numerical illustration from Section 3.3.3 that the welfare gains from group punishments are increasing in the parameter ρ. For a

given set of parameters, let ΔU denote the change in the value of the best equilibrium in our model relative to the value of the best equilibrium in the model where group punishments are not allowed, i.e.

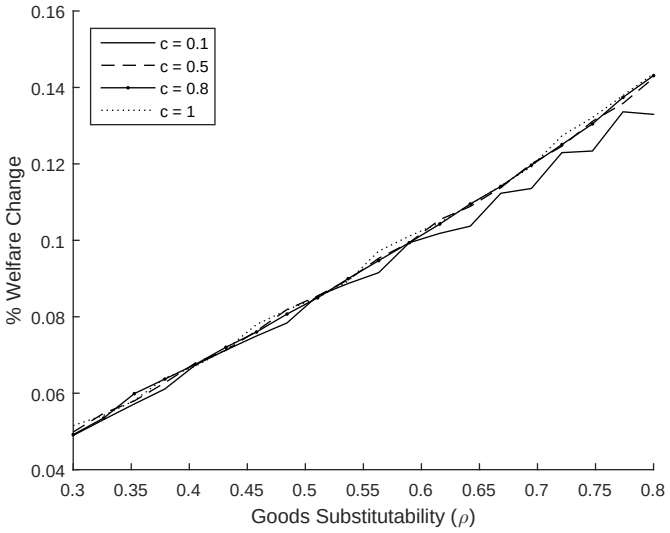
$$\Delta U \equiv \frac{u(\bar{q}) - u\left(\bar{q}^A\right)}{u\left(\bar{q}^A\right)}. \tag{D.11}$$

Proposition C.3. *Fix $\rho \in (0,1)$. For n sufficiently large, there exists a $\delta \in (0,1)$ and $\bar{\rho} > 0$ such that for all $\rho' \in (\rho, \bar{\rho})$, $d\Delta U(\rho')/d\rho' > 0$.*

We give here a sketch of our argument, and leave a formal proof to Appendix C.2.2.3. For a fixed level of the substitutability parameter ρ, we know that our model achieves the first-best level of output q^m at a lower level of the discount factor than the model where taxes are not allowed. This happens because, as showed in Proposition III.8, the threat of taxes always weakly enlarges the equilibrium set, and strictly enlarges the equilibrium set when producers are sufficiently patient. We denote by $\delta^{A*}(\rho)$ the threshold level of the discount factor at which the model where taxes are not allowed first achieves q^m as the most collusive level of output, and by $\delta^*(\rho)$ the level of the discount factor at which our model first achieves q^m as the most collusive level of output. Since $\delta^*(\rho) < \delta^{A*}(\rho)$, we can always find a discount factor δ^0 such that $\delta^*(\rho) < \delta^0 < \delta^{A*}(\rho)$. At δ^0 the model where taxes are allowed achieves q^m as the most collusive level of output, while the model where taxes are not allowed achieves a higher level of output (a lower value) than q^m. In the final step of the proof we argue that by continuity at this δ^0, if ρ increases by a sufficiently small amount to some $\bar{\rho}' > \rho$, the model where taxes are allowed still achieves q^m as the most collusive level of output. At δ^0, on the other hand, the most collusive level of output under $\bar{\rho}'$ is strictly greater than the most collusive level of output under ρ in the model where taxes are not allowed (from Lemma C.2). Therefore the increase in output (and decrease in value) relative to q^m (the most collusive level of output at δ^0, in the model where taxes are allowed) increases when ρ increases to $\bar{\rho}'$. Using the same argument, we prove that for all $\rho' \in (\rho, \bar{\rho})$, $d\Delta U(\rho')/d\rho' > 0$.

Different parametrizations of the model suggest that the results of Proposition C.3 hold for a wide range of the model's key parameters. As an example, Figure C.1 shows the percentage increase in welfare in the best equilibrium associated with group punishments for various values of the degree of substitutability ρ and the marginal cost of production, c. In this figure, we hold the discount factor fixed at a value of $\delta = 0.16$. This figure clearly shows that an increase in the degree of substitutability strictly raises the welfare gains associated with group punishments and that these welfare gains are not particularly sensitive to the marginal costs of production.

Figure C.1: Percentage increases in Welfare from Group Punishments



Percentage increase in welfare in best equilibrium resulting from the introduction of group punishments for various marginal costs of production c for a fixed discount factor ($\delta = 0.16$).

C.2 Definitions and Proofs

C.2.1 Definitions and Proofs from Sections 3.2 and 3.3

C.2.1.1 Repeated Game Definitions

Definition 1. *For any history $h^{wt} \in \mathcal{H}^w$ the continuation game is the infinitely-repeated game that begins in period t, following history h^{wt}. For any strategy profile $\sigma = (\{\sigma_i\}_{i=1}^n, \sigma_w)$, agent i's continuation strategy induced by h^{wt} is given by $\sigma_i(h^{wt}h^{ws})$ for all $h^{ws} \in \mathcal{H}^w$, where $h^{wt}h^{ws}$ is the concatenation of history h^{wt} followed by history h^{ws}. Similarly, the Principal continuation strategy induced by h^{wt} is given by $\sigma_w((h^{wt}h^{ws}), x(\sigma_1(h^{wt}h^{ws}), \sigma_2(h^{wt}h^{ws}), \ldots, \sigma_n(h^{wt}h^{ws})))$ for all $h^{ws} \in \mathcal{H}^w$.*

Definition 2. *A Perfect-Public Equilibrium is $\sigma = (\{\sigma_i\}_{i=1}^n, \sigma_w)$ such that, for all histories $h^{wt} \in \mathcal{H}^w$,*

$$U_i^t\left(h^{wt}, \sigma\right) \;\geq\; U_i^t\left(h^{wt}, (\tilde{\sigma}_i, \sigma_{-i}, \sigma_w)\right) \tag{D.12}$$

for all i, $\tilde{\sigma}_i$, and

$$U_w^t\left(h^{wt}, \sigma\right) \;\geq\; U_w^t\left(h^{wt}, (\{\sigma_i\}_{i=1}^n, \tilde{\sigma}_w)\right) \tag{D.13}$$

for all $\tilde{\sigma}_w$.

Definition 3. *A one-shot deviation for agent* i *from strategy* σ_i *is a strategy* $\tilde{\sigma}_i \neq \sigma_i$ *such that there exists a unique history* $\tilde{h}^{wt} \in \mathcal{H}^w$ *such that for all* $h^{ws} \neq \tilde{h}^{wt}$,

$$\sigma_i\left(h^{ws}\right) \;=\; \tilde{\sigma}_i\left(h^{ws}\right). \tag{D.14}$$

Similarly, a one-shot deviation for the Principal from strategy σ_w *is a strategy* $\tilde{\sigma}_w \neq \sigma_w$ *such that for all* $h^{wt} \in \mathcal{H}^w$ *there exists a level of the total outcome* $\tilde{x}_t$ *such that for all* $x_t \neq \tilde{x}_t$,

$$\sigma_w\left(h^{wt}, x_t\right) \;=\; \tilde{\sigma}_w\left(h^{wt}, x_t\right). \tag{D.15}$$

Definition 4. *A one-shot deviation* $\tilde{\sigma}_i$ *from the agent strategy* σ_i *is profitable if at history* $\tilde{h}^{wt}$ *for which* $\tilde{\sigma}_i\left(\tilde{h}^{wt}\right) \neq \sigma_i\left(\tilde{h}^{wt}\right)$,

$$U_i^t\left(\tilde{h}^{wt}, \left(\tilde{\sigma}_i, \sigma_{-i}, \sigma_w\right)\right) \;>\; U_i^t\left(\tilde{h}^{wt}, \sigma\right). \tag{D.16}$$

A one-shot deviation $\tilde{\sigma}_w$ *from the Principal strategy* σ_w *is profitable if for all* $h^{wt} \in \mathcal{H}^w$, *at the outcome level for which* $\tilde{\sigma}_w\left(\tilde{h}^{wt}, x_t\right) \neq \sigma_w\left(\tilde{h}^{wt}, x_t\right)$,

$$U_w^t\left(\tilde{h}^{wt}, \left(\{\sigma_i\}_{i=1}^n, \tilde{\sigma}_w\right)\right) \;>\; U_w^t\left(\tilde{h}^{wt}, \sigma\right). \tag{D.17}$$

C.2.1.2 Proof of Proposition III.3

If a profile is perfect-public, clearly there are no profitable one-shot deviations. Now suppose that the profile σ is not perfect-public. We want to show that there must be a profitable one-shot deviation. Since σ is not perfect-public, there exists a history $\tilde{h}^{wt}$, an agent i and a strategy $\tilde{\sigma}_i$ (the proof for the Principal follows the same steps) such that

$$U_i^t\left(\tilde{h}^{wt}, \sigma\right) \;<\; U_i^t\left(\tilde{h}^{wt}, \left(\tilde{\sigma}_i, \sigma_{-i}, \sigma_w\right)\right). \tag{D.18}$$

Let $\varepsilon = U_t^i\left(\tilde{h}^{wt}, \left(\tilde{\sigma}_i, \sigma_{-i}, \sigma_w\right)\right) - U_t^i\left(\tilde{h}^{wt}, \sigma\right)$. Let $m = \min_{i,q,\tau} u_i\left(q, \tau\right)$ and $M = \max_{i,q,\tau} u_i\left(q, \tau\right)$, with T large enough that $\delta^T\left(M - m\right) < \varepsilon/2$.[1]. Finally, for any agent i and history $h^{ws} \in \mathcal{H}^w$, let

$$u_i^s\left(\left(\tilde{h}^{wt} h^{ws}\right), \sigma\right) \;=\; u_i\left(\left\{\sigma_i\left(\tilde{h}^{wt} h^{ws}\right)\right\}_{i=1}^n, \sigma_w\left(\left(\tilde{h}^{wt} h^{ws}\right), x\left(\tilde{h}^{wt} h^{ws}\right)\right)\right), \tag{D.19}$$

[1]Note that $u_i\left(\cdot\right)$ is potentially unbounded below. Here we impose that m is an arbitrarily large negative number.

where $x\left(\tilde{h}^{wt}h^{ws}\right)$ is short-hand notation for $x\left(\sigma_1\left(\tilde{h}^{wt}h^{ws}\right),\sigma_2\left(\tilde{h}^{wt}h^{ws}\right),\ldots,\sigma_n\left(\tilde{h}^{wt}h^{ws}\right)\right)$, and denote by $\tilde{h}^{ws}$ the period-s history induced by $(\tilde{\sigma}_i,\sigma_{-i},\sigma_w)$. Then,

$$\begin{aligned}&(1-\delta)\left[\sum_{s=t}^{T-1}\delta^s u_i^s\left(\left(\tilde{h}^{wt}h^{ws}\right),\sigma\right)+\sum_{s=T}^{\infty}\delta^s u_i^s\left(\left(\tilde{h}^{wt}h^{ws}\right),\sigma\right)\right]\\&=(1-\delta)\left[\sum_{s=0}^{T-1}\delta^s u_i^s\left(\left(\tilde{h}^{wt}\tilde{h}^{ws}\right),(\tilde{\sigma}_i,\sigma_{-i},\sigma_w)\right)+\sum_{s=T}^{\infty}\delta^s u_i^s\left(\left(\tilde{h}^{wt}\tilde{h}^{ws}\right),(\tilde{\sigma}_i,\sigma_{-i},\sigma_w)\right)\right]-\varepsilon,\end{aligned} \tag{D.20}$$

so that

$$(1-\delta)\sum_{s=t}^{T-1}\delta^s u_i^s\left(\left(\tilde{h}^{wt}h^{ws}\right),\sigma\right) \;<\; (1-\delta)\sum_{s=0}^{T-1}\delta^s u_i^s\left(\left(\tilde{h}^{wt}\tilde{h}^{ws}\right),(\tilde{\sigma}_i,\sigma_{-i},\sigma_w)\right)-\frac{\varepsilon}{2} \tag{D.21}$$

Then the strategy $\hat{\sigma}_i$ such that

$$\hat{\sigma}_i\left(h^{ws}\right)=\begin{cases}\tilde{\sigma}_i\left(h^{ws}\right) & \text{if } s<T,\\ \sigma_i\left(h^{ws}\right) & \text{if } s\geq T,\end{cases} \tag{D.22}$$

is a profitable deviation from $\sigma_i\left(\tilde{h}^{wt}\right)$. Now let $\hat{h}^{w(T-1)}$ denote the period $T-1$ history induced by $(\hat{\sigma}_i,\sigma_{-i},\sigma_w)$. There are two possibilities. First, suppose

$$U_i^{T-1}\left(\left(\tilde{h}^{wt}\hat{h}^{w(T-1)}\right),\sigma\right) \;<\; U_i^{T-1}\left(\left(\tilde{h}^{wt}\hat{h}^{w(T-1)}\right),(\hat{\sigma}_i,\sigma_{-i},\sigma_w)\right). \tag{D.23}$$

Then, since $\hat{\sigma}_i$ agrees with σ_i in period T and after T, we have a profitable one-shot deviation after history $\tilde{h}^{wt}\hat{h}^{w(T-1)}$. Alternatively, suppose

$$U_i^{T-1}\left(\left(\tilde{h}^{wt}\hat{h}^{w(T-1)}\right),\sigma\right) \;\geq\; U_i^{T-1}\left(\left(\tilde{h}^{wt}\hat{h}^{w(T-1)}\right),(\hat{\sigma}_i,\sigma_{-i},\sigma_w)\right), \tag{D.24}$$

and construct the strategy

$$\bar{\sigma}_i\left(h^{ws}\right) \;=\; \begin{cases}\hat{\sigma}_i\left(h^{ws}\right) & \text{if } s<T-1,\\ \sigma_i\left(h^{ws}\right) & \text{if } s\geq T-1.\end{cases} \tag{D.25}$$

Since

$$\begin{aligned} U_i^{T-2}\left(\left(\tilde{h}^{wt}\hat{h}^{w(T-2)}\right),(\hat{\sigma}_i,\sigma_{-i},\sigma_w)\right) &= (1-\delta)\,u_i^{T-2}\left(\left(\tilde{h}^{wt}\hat{h}^{w(T-2)}\right),(\hat{\sigma}_i,\sigma_{-i},\sigma_w)\right) \\ &+ \delta U_i^{T-1}\left(\left(\tilde{h}^{wt}\hat{h}^{w(T-1)}\right),(\hat{\sigma}_i,\sigma_{-i},\sigma_w)\right) \qquad \text{(D.26)} \\ &\leq (1-\delta)\,u_i^{T-2}\left(\left(\tilde{h}^{wt}\hat{h}^{w(T-2)}\right),(\hat{\sigma}_i,\sigma_{-i},\sigma_w)\right) \\ &+ \delta U_i^{T-1}\left(\left(\tilde{h}^{wt}\hat{h}^{w(T-1)}\right),\sigma\right) \qquad \text{(D.27)} \\ &= U_i^{T-2}\left(\left(\tilde{h}^{wt}\hat{h}^{w(T-2)}\right),(\bar{\sigma}_i,\sigma_{-i},\sigma_w)\right), \qquad \text{(D.28)} \end{aligned}$$

then

$$U_i^t\left(\tilde{h}^{wt},(\hat{\sigma}_i,\sigma_{-i},\sigma_w)\right) \leq U_i^t\left(\tilde{h}^{wt},(\bar{\sigma}_i,\sigma_{-i},\sigma_w)\right), \qquad \text{(D.29)}$$

and $\bar{\sigma}_i$ is a profitable deviation at $\tilde{h}^{wt}$ that only differs from σ_i in the first $T-1$ periods. Proceeding in this way, we find a profitable one-shot deviation.

C.2.1.3 Proof of Proposition III.6

We need only prove that for each $v \in [\underline{v},\bar{v}]$, there exists a perfect-public equilibrium strategy which attains the value v. To construct such strategy, we start from the set of perfect-public equilibrium strategies of the game where the Principal is not allowed to impose group punishments, $\left[\underline{v}^A,\bar{v}^A\right]$. We know from *Abreu* (1986) that any equilibrium value v^0 such that $v^0 \in \left[\underline{v}^A,\bar{v}^A\right]$ can be achieved with a perfect-public equilibrium strategy σ^0. Under σ^0, the Principal never imposes group punishments and agents exert effort a^0 such that $u\left(a^0\right)=v^0$ on path, and punish deviations by both Principal and agents by reversion to the worst (carrot-and-stick) perfect-public equilibrium with value $\underline{v}^A$. Therefore, we focus on characterizing the equilibrium strategies for the cases in which $\left[\underline{v}^A,\bar{v}^A\right] \subset [\underline{v},\bar{v}]$.

Consider a new strategy σ^1. Define by $\bar{a}^A$ the carrot output in the model where group punishments are not allowed. Under σ^1, for some $\varepsilon^1>0$ agents choose $a^1=\bar{a}^A+\varepsilon^1$ as long as the aggregate outcome x^1 is such that $x^1=x\left(a^1\right)$, and the Principal never imposes punishments. Suppose that an agent deviates to some a', such that the observed aggregate outcome is $x^1=x\left(a',a^1\right)$. In this case, the Principal imposes an arbitrarily small punishment $\tau^1\left(x^1\right)>0$ such that the punishment is feasible. That is, such that $v^1\left(a',a^1,\tau^1\left(x^1\right)\right)\in[\underline{v},\bar{v}]$, where

$$v^1\left(a',a^1,\tau^1\left(x^1\right)\right) \equiv \frac{1-\delta}{\delta}\frac{1}{n}\left[w\left(a',a^1,0\right)-w\left(a',a^1,\tau^1\left(x^1\right)\right)\right]. \qquad \text{(D.30)}$$

If an agent deviates and the Principal implements the prescribed punishment, then agents follow the strategy $\sigma^1\left(v^1\left(a', a^1, \tau^1\left(x^1\right)\right)\right)$. Therefore, the continuation value promised to agents when one of the agents deviates and the Principal imposes $\tau^1\left(x^1\right)$ can be achieved with a perfect-public equilibrium strategy. Conversely, deviations by agents followed by deviations by the Principal are punished by the worst perfect-public equilibrium strategy $\sigma^1\left(\underline{v}^A\right)$. Clearly, this strategy is a perfect-public equilibrium. Moreover, it achieves a value $u\left(a^1\right) \equiv \bar{v}^1 > \bar{v}^A$.

Next, note that reversion to the perfect-public equilibrium $\bar{v}^1 > \bar{v}^A$ allows to construct a new carrot-and-stick strategy in which agents contribute an effort level $\tilde{a}^1 < \tilde{a}^A$ for one period and then revert to $\bar{v}^1$, with deviations from the prescription causing the prescription to be repeated. Moreover, note that this new carrot-and-stick strategy has value $\underline{v}^1 < \underline{v}^A$. Hence, for any value $v^1 \in [\underline{v}^1, \bar{v}^1]$, we can find a perfect-public equilibrium strategy σ^1 such that $u\left(\sigma^1\right) = v^1$.

Now take some $k \geq 2$ and set $\left[\underline{v}^k, \bar{v}^k\right]$ such that $[\underline{v}^1, \bar{v}^1] \subset \left[\underline{v}^k, \bar{v}^k\right] \subset [\underline{v}, \bar{v}]$, and assume that for any $v^k \in \left[\underline{v}^k, \bar{v}^k\right]$ we can construct a perfect-public equilibrium strategy σ^k such that $u\left(\sigma^k\right) = v^k$. Denote by $\bar{a}^k$ the effort level with value $\bar{v}^k$, and construct a new strategy σ^{k+1}. Under σ^{k+1}, for some $\varepsilon^{k+1} > 0$ agents produce $a^{k+1} = \bar{a}^k + \varepsilon^{k+1}$ as long as the observed aggregate outcome x^{k+1} is such that $x^{k+1} = x\left(a^{k+1}\right)$, and the Principal never imposes punishments. Suppose that an agent deviates to some a', such that the observed aggregate outcome is $x^{k+1} = x\left(a', a^{k+1}\right)$. In this case, the Principal imposes a punishment $\tau^{k+1}\left(x^{k+1}\right) > 0$ such that the punishment is feasible. That is, such that $v^{k+1}\left(a', a^{k+1}, \tau^{k+1}\left(x^{k+1}\right)\right) \in [\underline{v}, \bar{v}]$, where

$$v^{k+1}\left(a', a^{k+1}, \tau^{k+1}\left(x^{k+1}\right)\right) \equiv \frac{1-\delta}{\delta}\frac{1}{n}\left[w\left(a', a^{k+1}, 0\right) - w\left(a', a^{k+1}, \tau^{k+1}\left(x^{k+1}\right)\right)\right] \quad \text{(D.31)}$$

Note that since $\bar{v}^k > \bar{v}^1$, the range of punishments that can be sustained is larger than $[0, \sup_{x^1} \tau^1\left(x^1\right)]$. If an agent deviates and the Principal implements the prescribed tax, then agents follow the strategy $\sigma^{k+1}\left(v^{k+1}\left(a', a^{k+1}, \tau^{k+1}\left(x^{k+1}\right)\right)\right)$. Therefore, the continuation value promised to agents when one of the agents deviates and the Principal imposes $\tau^{k+1}\left(x^{k+1}\right)$ can be achieved with a perfect-public equilibrium strategy. Conversely, deviations by agents followed by deviations by the Principal are punished by the worst perfect-public equilibrium strategy $\sigma^{k+1}\left(\underline{v}^k\right)$. Clearly, this strategy is a perfect-public equilibrium. Moreover, it achieves a value $u\left(a^{k+1}\right) \equiv \bar{v}^{k+1} > \bar{v}^k$. Next, note that reversion to the perfect-public equilibrium $\bar{v}^{k+1} > \bar{v}^k$ allows to construct a new carrot-and-stick strategy in which agents exert an effort level $\tilde{a}^{k+1} > \tilde{a}^k$ for one period and then revert to $\bar{v}^{k+1}$, with deviations from the prescription causing the prescription to be repeated. Moreover, note that this new carrot-and-stick strategy has value $\underline{v}^{k+1} < \underline{v}^k$. Hence, for any value $v^{k+1} \in \left[\underline{v}^{k+1}, \bar{v}^{k+1}\right]$, we

can find a perfect-public equilibrium strategy σ^{k+1} such that $u\left(\sigma^{k+1}\right) = v^{k+1}$. The proof is completed by induction.

C.2.1.4 Proof of Proposition III.7

Suppose $\sigma\left(\left(\tilde{a},\bar{a}\right),(0,0)\right)$ is an optimal carrot-and-stick punishment. Recalling from Proposition III.5 that $\tilde{a} \leq a^N$, the requirement that producers do not deviate from the stick and carrot outputs $\tilde{a}$ and $\bar{a}$ are, respectively:

$$\begin{aligned}
(1-\delta)\,u\left(\tilde{a},0\right)+\delta u\left(\bar{a},0\right) &\geq (1-\delta)\,\hat{g}\left(\tilde{a},0\right)+\delta\left(1-\delta\right)u\left(\tilde{a},0\right)+\delta^2 u\left(\bar{a},0\right), &\text{(D.32)}\\
u\left(\bar{a},0\right) &\geq (1-\delta)\,\hat{g}\left(\bar{a},\tau\left(\cdot\right)\right)+\delta\left(1-\delta\right)u\left(\tilde{a},0\right)+\delta^2 u\left(\bar{a},0\right). &\text{(D.33)}
\end{aligned}$$

Rearranging these inequalities, we get

$$\begin{aligned}
\hat{g}\left(\tilde{a},0\right) &\leq (1-\delta)\,u\left(\tilde{a},0\right)+\delta u\left(\bar{a},0\right)=\underline{v}, &\text{(D.34)}\\
\hat{g}\left(\bar{a},\tau\left(\cdot\right)\right) &\leq u\left(\bar{a},0\right)+\delta\left(u\left(\bar{a},0\right)-u\left(\tilde{a},0\right)\right). &\text{(D.35)}
\end{aligned}$$

If (D.34) holds strictly, we can decrease $\tilde{a}$ and hence reduce $u\left(\tilde{a},0\right)$ while preserving (D.35). But this yields a lower punishment value than the infimum $\underline{v}$, a contradiction. Hence (D.34) holds with equality. Now suppose that if $\bar{a} < a^*$, (D.35) holds as a strict inequality. Then we can simultaneously decrease $\tilde{a}$ by a small amount (therefore not violating (D.35)) and increase $\bar{a}$ to preserve (D.34). But then since $\hat{g}\left(\tilde{a},0\right)$ is increasing in $\tilde{a}$ and (D.34), we also found a lower punishment value than the infimum, again a contradiction.

C.2.1.5 Proof of Lemma III.10

First, note that

$$\begin{aligned}
\hat{g}\left(q,\tau\left(\cdot\right)\right) &\geq \max_{q'} g\left(q',q,\tau\left(q'+\left(n-1\right)q\right)\right) &\text{(D.36)}\\
&\geq g\left(\frac{1-\left(n-1\right)q-c}{2},q,\tau\left(\frac{1-\left(n-1\right)q-c}{2}+\left(n-1\right)q\right)\right). &\text{(D.37)}
\end{aligned}$$

Moreover, note that

$$\begin{aligned}
\frac{\partial g\left(q',q,\tau\right)}{\partial\tau} &= -q'+\frac{1}{n}\left(q'+\left(n-1\right)q\right) &\text{(D.38)}\\
&= \frac{n-1}{n}\left(q-q'\right), &\text{(D.39)}
\end{aligned}$$

so that $\partial g\left(q, q', \tau\right)/\partial\tau \geq 0$ if and only if $q' \leq q$. Finally, for $q \geq q^N$ if we choose the deviation

$$q' = \frac{1-(n-1)\,q-c}{2}, \tag{D.40}$$

then it must be that $q' \leq q$, since

$$\frac{1-(n-1)\,q-c}{2} \leq q \iff q^N \leq q. \tag{D.41}$$

Hence, for $q \geq q^N$,

$$\begin{aligned}
\hat{g}\left(q, \tau\left(\cdot\right)\right) &\geq g\left(\frac{1-(n-1)\,q-c}{2}, q, \tau\left(\frac{1-(n-1)\,q-c}{2}+(n-1)\,q\right)\right) && \text{(D.42)}\\
&\geq g\left(\frac{1-(n-1)\,q-c}{2}, q, 0\right) && \text{(D.43)}\\
&= \hat{g}\left(q, 0\right). && \text{(D.44)}
\end{aligned}$$

C.2.2 Proofs from Appendix C.1

C.2.2.1 Proof of Lemma C.1

The proof is a straightforward extension of the proof found in Appendix C.2.1.5. In absence of a closed-form for the optimal deviation for the model where taxes are not allowed, the only additional step required to complete the proof is to show that for $q \geq q^N$ if we choose the deviation

$$q' = \hat{q}\left(q\right), \tag{D.45}$$

then $q' \leq q$. We prove this by showing that $\hat{q}\left(q\right)$ is a smooth function that only intersects the 45 degree line at zero and q^N, and that for some $q > q^N$, $\hat{q}\left(q\right) < q$. First, note that $\hat{q}\left(0\right) = 0$, and that by definition of Nash equilibrium for $q > 0$, $\hat{q}\left(q\right) = q$ if and only if $q = q^N$. Moreover, note that $\hat{q}\left(q\right)$ is smooth, since the problem is smooth and $\hat{q}\left(q\right)$ is the implicit function that generates from the first order conditions determining the most profitable deviation from q. Finally, note that for some $q > q^N$, $\hat{q}\left(q\right) < q$. To show this, consider any $q > q^0$, where q^0 is the minimum $q > 0$ such that $\hat{q}\left(q\right) = 0$ if and only if $\hat{g}\left(q, \tau = 0\right) = 0$. This level of q exists (as q goes to infinity, the price is driven to zero and the most profitable deviation is not to produce and avoid the associated cost) and we can always find it large enough such that $q > q^N$.

C.2.2.2 Proof of Lemma C.2

Since in what follows we focus on a model where taxes are not allowed, for notational convenience we drop functional dependencies on taxes (e.g. we denote $u(q,0)$ by $u(q)$). We similarly drop the superscript "A" which we use to compare the model where taxes are allowed to the model where taxes are not allowed. To show that for sufficiently large n, $d\bar{q}/d\rho > 0$, we analyze the two equations that characterize the carrot and the stick output in the model where taxes are not allowed, i.e.

$$g\left(\tilde{q}\left(\rho\right);\rho\right) = \frac{\alpha}{n} - (1-\delta)c\tilde{q}\left(\rho\right) - \delta c\bar{q}\left(\rho\right), \tag{D.46}$$

$$g\left(\bar{q}\left(\rho\right);\rho\right) = \frac{\alpha}{n} - (1+\delta)c\bar{q}(\rho) + \delta c\tilde{q}\left(\rho\right). \tag{D.47}$$

Totally differentiating these two expressions, we obtain

$$g_q\left(\tilde{q}\left(\rho\right);\rho\right)\frac{d\tilde{q}(\rho)}{d\rho} + g_\rho\left(\tilde{q}\left(\rho\right);\rho\right) = -(1-\delta)c\frac{d\tilde{q}(\rho)}{d\rho} - \delta c\frac{d\bar{q}\left(\rho\right)}{d\rho} \tag{D.48}$$

$$g_q\left(\bar{q}\left(\rho\right);\rho\right)\frac{d\bar{q}(\rho)}{d\rho} + g_\rho\left(\bar{q}\left(\rho\right);\rho\right) = -(1+\delta)c\frac{d\bar{q}\left(\rho\right)}{d\rho} + \delta c\frac{d\tilde{q}\left(\rho\right)}{d\rho} \tag{D.49}$$

We solve for $d\tilde{q}/d\rho$ from the first equation and substitute into the second to obtain a form for $d\bar{q}/d\rho$. We have

$$\left[g_q\left(\tilde{q}\left(\rho\right);\rho\right) + (1-\delta)\,c\right]\frac{d\tilde{q}(\rho)}{d\rho} = -g_\rho\left(\tilde{q}\left(\rho\right);\rho\right) - \delta c\frac{d\bar{q}\left(\rho\right)}{d\rho}, \tag{D.50}$$

so that

$$\left[g_q\left(\bar{q}\left(\rho\right);\rho\right) + (1+\delta)c\right]\frac{d\bar{q}(\rho)}{d\rho} = -g_\rho\left(\bar{q}\left(\rho\right);\rho\right) + \delta c\left[\frac{-g_\rho\left(\tilde{q}\left(\rho\right);\rho\right) - \delta c\frac{d\bar{q}(\rho)}{d\rho}}{g_q\left(\tilde{q}\left(\rho\right);\rho\right) + (1-\delta)\,c}\right] \tag{D.51}$$

To be able to determine the sign of $d\bar{q}/d\rho$, we determine the sign of the derivatives $g_q\left(\hat{q};\rho\right)$ and $g_q\left(\hat{q};\rho\right)$. First, for any $\hat{q}$ we denote the most profitable deviation from $\hat{q}$ as a function of $\hat{q}$ and ρ as $q^*\left(\hat{q},\rho\right)$. From the optimality conditions, we know that any $q^*\left(\hat{q},\rho\right)$ satisfies

$$\frac{\alpha\rho\left(n-1\right)\hat{q}^\rho\left(q^*\right)^{-\rho-1}}{\left[1+\left(n-1\right)\hat{q}^\rho\left(q^*\right)^{-\rho}\right]^2} = c. \tag{D.52}$$

Next, note that the payoff from the best response satisfies

$$\begin{aligned} g_q(\hat{q};\rho) &= \frac{d}{d\hat{q}}\left[\frac{\alpha}{1+(n-1)\,\hat{q}^{\rho}q^*(\hat{q},\rho)^{-\rho}} - q^*(\hat{q},\rho)\,c\right] && \text{(D.53)}\\ &= \frac{-\alpha}{\left[1+(n-1)\,\hat{q}^{\rho}q^*(\hat{q},\rho)^{-\rho}\right]^2}\left[(n-1)\,\rho\hat{q}^{\rho-1}q^*(\hat{q},\rho)^{-\rho}\right. \\ &- \left.(n-1)\,\rho\hat{q}^{\rho}q^*(\hat{q},\rho)^{-\rho-1}\,q_q^*(\hat{q},\rho)\right] - cq_q^*(\hat{q},\rho) && \text{(D.54)}\\ &= \frac{-\alpha\rho\,(n-1)\,\hat{q}^{\rho-1}q^*(\hat{q},\rho)^{-\rho}}{\left[1+(n-1)\,\hat{q}^{\rho}q^*(\hat{q},\rho)^{-\rho}\right]^2} + \left[\frac{\alpha\,(n-1)\,\rho\hat{q}^{\rho}q^*(\hat{q},\rho)^{-\rho-1}}{\left[1+(n-1)\,\hat{q}^{\rho}q^*(\hat{q},\rho)^{-\rho}\right]^2} - c\right]q_q^*(\hat{q},\rho) && \text{(D.55)}\\ &= \frac{-\alpha\rho\,(n-1)\,\hat{q}^{\rho}\,(q^*)^{-\rho}}{\hat{q}\left[1+(n-1)\,\hat{q}^{\rho}\,(q^*)^{-\rho}\right]^2}, && \text{(D.56)} \end{aligned}$$

where the last equality follows from optimality of q^*. Note also that using optimality of q^*, we may write $g_q(\hat{q};\rho)$ as

$$g_q(\hat{q};\rho) = \frac{-\alpha\rho\,(n-1)\,\hat{q}^{\rho}\,(q^*)^{-\rho}}{\hat{q}\left[1+(n-1)\,\hat{q}^{\rho}\,(q^*)^{-\rho}\right]^2} = -\frac{cq^*(\hat{q},\rho)}{\hat{q}}. \qquad \text{(D.57)}$$

Since if $\hat{q} < q^N$ then $q^*(\hat{q},\rho) \geq \hat{q}$ and if $\hat{q} \geq q^N$ then $q^*(\hat{q},\rho) \leq q^N$, this implies

$$\begin{aligned} g_q(\hat{q};\rho) &\leq -c \text{ if } \hat{q} < q^N, && \text{(D.58)}\\ g_q(\hat{q};\rho) &\geq -c \text{ if } \hat{q} \geq q^N. && \text{(D.59)} \end{aligned}$$

Similarly, note that

$$\begin{aligned} g_\rho(\hat{q};\rho) &= \frac{d}{d\rho}\left[\frac{\alpha}{1+(n-1)\,\hat{q}^{\rho}q^*(\hat{q},\rho)^{-\rho}} - q^*(\hat{q},\rho)\,c\right] && \text{(D.60)}\\ &= \frac{-\alpha\,(n-1)\,q^*(\hat{q},\rho)^{-\rho}\,\hat{q}^{\rho}}{\left[1+(n-1)\,\hat{q}^{\rho}q^*(\hat{q},\rho)^{-\rho}\right]^2}\left[\log\hat{q} - \log q^*(\hat{q},\rho)\right], && \text{(D.61)} \end{aligned}$$

so that we can write

$$g_\rho(\hat{q};\rho) = -\frac{c}{\rho}q^*(\hat{q},\rho)\left[\log\hat{q} - \log q^*(\hat{q},\rho)\right]. \qquad \text{(D.62)}$$

We then have

$$\begin{aligned} g_\rho(\hat{q};\rho) &\geq 0 \text{ if } \hat{q} < q^N, && \text{(D.63)}\\ g_\rho(\hat{q};\rho) &\leq 0 \text{ if } \hat{q} \geq q^N. && \text{(D.64)} \end{aligned}$$

Next, substituting (D.57) and (D.62) into (D.51), we obtain

$$\begin{aligned}\frac{d\bar{q}(\rho)}{d\rho}\left[g_q\left(\bar{q}\left(\rho\right);\rho\right)+(1+\delta)c+\frac{\delta^2c^2}{g_q\left(\tilde{q}\left(\rho\right);\rho\right)+(1-\delta)\,c}\right] &= -g_\rho\left(\bar{q}\left(\rho\right);\rho\right)\\ &- \frac{\delta c g_\rho\left(\tilde{q}\left(\rho\right);\rho\right)}{g_q\left(\tilde{q}\left(\rho\right);\rho\right)+(1-\delta)\,c}\end{aligned} \tag{D.65}$$

Simplifying and using short-hand notation, we have

$$\bar{q}_\rho\left[-\frac{\bar{q}^*}{\bar{q}}+(1+\delta)+\frac{\delta^2}{-\frac{\tilde{q}^*}{\tilde{q}}+(1-\delta)}\right] = \frac{1}{\rho}\bar{q}^*\log\left(\frac{\bar{q}}{\bar{q}^*}\right)+\frac{\delta\frac{1}{\rho}\tilde{q}^*\log\left(\frac{\tilde{q}}{\tilde{q}^*}\right)}{-\frac{\tilde{q}^*}{\tilde{q}}+(1-\delta)} \tag{D.66}$$

$$\bar{q}_\rho\left[1-\frac{\bar{q}^*}{\bar{q}}+\frac{\delta\left[1-\frac{\tilde{q}^*}{\tilde{q}}\right]}{1-\frac{\tilde{q}^*}{\tilde{q}}-\delta}\right] = \frac{1}{\rho}\bar{q}^*\log\left(\frac{\bar{q}}{\bar{q}^*}\right)+\frac{\delta\frac{1}{\rho}\tilde{q}^*\log\left(\frac{\tilde{q}}{\tilde{q}^*}\right)}{1-\frac{\tilde{q}^*}{\tilde{q}}-\delta} \tag{D.67}$$

$$\bar{q}_\rho\left[\frac{\bar{q}-\bar{q}^*}{\bar{q}}+\frac{\delta\left[\tilde{q}-\tilde{q}^*\right]}{(1-\delta)\tilde{q}-\tilde{q}^*}\right] = \frac{1}{\rho}\bar{q}^*\log\left(\frac{\bar{q}}{\bar{q}^*}\right)+\frac{\delta\frac{1}{\rho}\tilde{q}\tilde{q}^*\log\left(\frac{\tilde{q}}{\tilde{q}^*}\right)}{(1-\delta)\tilde{q}-\tilde{q}^*}. \tag{D.68}$$

Since $\bar{q}\leq\bar{q}^*$ and $\tilde{q}\geq\tilde{q}^*$, if

$$(1-\delta)\tilde{q}-\tilde{q}^*\leq 0, \tag{D.69}$$

then each term in brackets on the left-hand side and each term on the right-hand side of (D.68) are negative. This implies $\bar{q}_\rho\geq 0$. Hence, $(1-\delta)\tilde{q}-\tilde{q}^*\leq 0$ is a sufficient condition for $\bar{q}_\rho$ to be positive. To show this, we use the expression for $g(\tilde{q},\rho)$ and $g(\bar{q},\rho)$ in (D.46) and (D.47):

$$\frac{\alpha}{1+(n-1)\left(\frac{\tilde{q}}{\tilde{q}^*}\right)^\rho}-c\tilde{q}^* = \frac{\alpha}{n}-(1-\delta)c\tilde{q}-\delta c\bar{q}, \tag{D.70}$$

$$\frac{\alpha}{1+(n-1)\left(\frac{\bar{q}}{\bar{q}^*}\right)^\rho}-c\bar{q}^* = \frac{\alpha}{n}-(1+\delta)c\bar{q}+\delta c\tilde{q}. \tag{D.71}$$

Equation (D.71) implies

$$(1+\delta)c\bar{q}-c\bar{q}^* = \frac{\alpha}{n}-\frac{\alpha}{1+(n-1)\left(\frac{\bar{q}}{\bar{q}^*}\right)^\rho}+\delta c\tilde{q} \tag{D.72}$$

$$\leq \delta c\tilde{q}, \tag{D.73}$$

which substituted into (D.70) yields

$$(1-\delta)c\tilde{q} - c\tilde{q}^* = \frac{\alpha}{n} - \frac{\alpha}{1+(n-1)\left(\frac{\tilde{q}}{\bar{q}^*}\right)^\rho} - \delta c\bar{q} \tag{D.74}$$

$$= (1+\delta)c\bar{q} - c\bar{q}^* + \frac{\alpha}{1+(n-1)\left(\frac{\bar{q}}{\bar{q}^*}\right)^\rho} - \frac{\alpha}{1+(n-1)\left(\frac{\tilde{q}}{\tilde{q}^*}\right)^\rho} - \delta c\left(\tilde{q}+\bar{q}\right) \tag{D.75}$$

$$\leq \delta c\tilde{q} + \frac{\alpha}{1+(n-1)\left(\frac{\bar{q}}{\bar{q}^*}\right)^\rho} - \frac{\alpha}{1+(n-1)\left(\frac{\tilde{q}}{\tilde{q}^*}\right)^\rho} - \delta c\left(\tilde{q}+\bar{q}\right). \tag{D.76}$$

Then, we have

$$(1-\delta)c\tilde{q} - c\tilde{q}^* \leq -\delta c\bar{q} + \frac{\alpha}{1+(n-1)\left(\frac{\bar{q}}{\bar{q}^*}\right)^\rho} - \frac{\alpha}{1+(n-1)\left(\frac{\tilde{q}}{\tilde{q}^*}\right)^\rho}. \tag{D.77}$$

For n sufficiently large, $\bar{q}^*$ converges to $\bar{q}$ and $\tilde{q}^*$ converges to $\tilde{q}$. Hence, for n sufficiently large the right-hand side is less than or equal to zero and the needed condition is verified.

C.2.2.3 Proof of Proposition C.3

Fix $\rho \in (0,1)$. We know that there exist a unique $\delta^{A*}(\rho)$ in the model where taxes are not allowed such that $\bar{q}^A(\rho) = q^m$. This $\delta^{A*}(\rho)$ simultaneously solves

$$\hat{g}\left(q^m, 0\right) = \left(1+\delta^{A*}(\rho)\right) u\left(q^m\right) - \delta^{A*}(\rho)\, u\left(\tilde{q}^A(\rho)\right), \tag{D.78}$$

$$\hat{g}\left(\tilde{q}^A(\rho), 0\right) = \left(1-\delta^{A*}(\rho)\right) u\left(\tilde{q}^A(\rho)\right) + \delta^{A*}(\rho)\, u\left(q^m\right), \tag{D.79}$$

and represents the threshold level of the discount factor for which the model where taxes are not allowed achieves the first-best level of output q^m. Similarly, for the same ρ we know that there exists a unique $\delta^*(\rho)$ in the model where taxes are allowed such that $\bar{q}(\rho) = q^m$, which simultaneously solves

$$\hat{g}\left(q^m, \tau(\cdot)\right) = \left(1+\delta^*(\rho)\right) u\left(q^m\right) - \delta^*(\rho)\, u\left(\tilde{q}(\rho)\right), \tag{D.80}$$

$$\hat{g}\left(\tilde{q}(\rho), 0\right) = \left(1-\delta^*(\rho)\right) u\left(\tilde{q}(\rho)\right) + \delta^*(\rho)\, u\left(q^m\right). \tag{D.81}$$

Next, note that since i) for any level of the discount factor we have $\left[\underline{v}^A; \bar{v}^A\right] \subseteq [\underline{v}; \bar{v}]$, and ii) for $\bar{q}^A > q^m$ if $\bar{q}$ is sustained by a positive tax threat (for some $q' \neq \bar{q}$, $\tau\left(q' + (n-1)\bar{q}\right) > 0$) then $\bar{q}^A > \bar{q} \geq q^m$, then $\delta^*(\rho) < \delta^{A*}(\rho)$ (i.e. the model where taxes are allowed achieves the first best level of output q^m at a lower value of the discount factor than the model where taxes are not allowed). Next, let δ^0 be such that $\delta^*(\rho) < \delta^0 < \delta^{A*}(\rho)$. Note that at δ^0,

$\bar{q}(\rho) = q^m$ and $\bar{q}^A(\rho) > q^m$. Now let $\bar{\rho}' > \rho$, and let $\delta^*(\bar{\rho}')$ in the model where taxes are allowed be such that $\bar{q}(\bar{\rho}') = q^m$, which solves

$$\hat{g}(q^m, \tau(\cdot)) = (1+\delta^*(\bar{\rho}'))\, u(q^m) - \delta^*(\rho)\, u(\tilde{q}(\rho')) \tag{D.82}$$

$$\hat{g}(\tilde{q}(\bar{\rho}'), 0) = (1-\delta^*(\bar{\rho}'))\, u(\tilde{q}(\bar{\rho}')) + \delta^*(\rho)\, u(q^m). \tag{D.83}$$

By continuity we know that we can always choose $\bar{\rho}'$ small enough such that $\delta^*(\bar{\rho}') < \delta^0$. Therefore, in the model where taxes are allowed $\bar{q}(\bar{\rho}') = \bar{q}(\rho) = q^m$. Moreover, since from Lemma C.2 we know that for n sufficiently large $d\bar{q}^A/d\rho > 0$, then $\bar{q}^A(\bar{\rho}') > \bar{q}^A(\rho)$. Hence, at δ^0

$$\frac{u(\bar{q}(\bar{\rho}')) - u\left(\bar{q}^A(\bar{\rho}')\right)}{u(\bar{q}^A(\bar{\rho}'))} > \frac{u(\bar{q}(\rho)) - u\left(\bar{q}^A(\rho)\right)}{u(\bar{q}^A(\rho))}. \tag{D.84}$$

Finally, following the same argument we have that for all $\rho' \in (\rho, \bar{\rho}')$, $d\Delta U(\rho')/d\rho' > 0$.

C.3 Computational Algorithm

In this Appendix, we describe the computational algorithm for our numerical results in Section 3.3. Define $\hat{q} \equiv \arg\max_{q'} \hat{g}(q', \bar{q}, \tau(q' + (n-1)\bar{q}))$. For each level of the discount factor δ, we aim to find $\tilde{q}$, $\bar{q}$, $\hat{q}$ and τ that solve the following system of equations:

$$\hat{g}(\tilde{q}, 0) = (1-\delta)\, u(\tilde{q}, 0) + \delta\mu(\bar{q}, 0), \tag{D.85}$$

$$\hat{g}(\hat{q}, \tau(\cdot)) \leq u(\bar{q}, 0) + \delta(u(\bar{q}, 0) - u(\tilde{q}, 0)), \tag{D.86}$$

$$u(\bar{q}, 0) \geq \frac{1-\delta}{\delta}\frac{1}{n}\left[w(\hat{q} + (n-1)\bar{q}, 0) - w(\hat{q} + (n-1)\bar{q}, \tau(\hat{q} + (n-1)\bar{q}))\right] + \hat{g}(\tilde{q}, 0). \tag{D.87}$$

From Proposition III.7, Equation (D.86) holds with equality only when $\bar{q} > q^m$ and is slack when $\bar{q} = q^m$. The algorithm works as follows:

1. For each level of the discount factor δ, we know $\tau \in \left[0, 1-(n-1)q^N - c\right]$. Start with $\hat{\tau} = 1-(n-1)q^N - c$.

   (a) Check if q^m can be supported:

      i. Set $\bar{q} = q^m$. Solve (D.85) for $\tilde{q}$.

      ii. Obtain $\hat{q} = \arg\max_{q' \in [q^m, q^N]} \hat{g}(q', \bar{q}, \hat{\tau})$. We do this by searching for $\hat{q}$ over a fine grid for q'. Evaluate $\hat{g}(\bar{q}, \hat{\tau})$.

iii. Check if the resulting values for $\tilde{q}$ and $\bar{q}$ satisfy (D.86) (with inequality) and (D.87). If so, the algorithm is finished.

(b) If either (D.86) or (D.87) is not satisfied (q^m cannot be supported), jointly solve for $\bar{q}$ and $\tilde{q}$. We do this using a nested bisection algorithm to solve (D.85) and (D.86) with equality (also solving for $\hat{q}$ as before).

i. The nested bisection algorithm proceeds as follows. The outer bisection algorithm searches for $\tilde{q} \in [\tilde{q}_\ell, \tilde{q}_h]$. The inner bisection algorithm solves for the corresponding $\bar{q}$.

ii. At each iteration of the double bisection algorithm, check whether (D.85)-(D.87) are all satisfied.

2. If (D.85) and (D.87) are satisfied, we are done. If not decrease $\hat{\tau}$ by a small amount and return to step 1.

BIBLIOGRAPHY

BIBLIOGRAPHY

Abreu, D. (1986), Extremal equilibria of oligopolistic supergames, *Journal of Economic Theory*, *39*(1), 191–225.

Acemoglu, D., and A. Wolitzky (2015), Sustaining cooperation: Community enforcement vs. specialized enforcement, *Tech. rep.*, National Bureau of Economic Research.

Acs, Z. J., R. Morck, and B. Yeung (1999), Productivity growth and firm size distribution, *Z. J. Acs, B. Carlsson, and C. Karlsson (eds.), Entrepreneurship, Small and Medium-Sized Enterprises and the Macroeconomy*, pp. 367–396.

Alchian, A. A., and H. Demsetz (1972), Production, information costs, and economic organization, *The American Economic Review*, *62*(5), 777–795.

Aldashev, G., and G. Zanarone (2017), Endogenous enforcement institutions, *Journal of Development Economics*, *128*, 49–64.

Allen, J., R. Clark, and J.-F. Houde (2008), Market structure and the diffusion of electronic banking, in *Federal Reserve Bank of Boston Workshop on Consumer Behavior and Payment Choice. Available at http://www. bos. frb. org/economic/eprg/conferences/payments2008/allen_clark_houde. pdf.*

Álvarez, R., and S. Claro (2009), David versus goliath: The impact of chinese competition on developing countries, *World Development*, *37*(3), 560–571.

Anderson, S. P., and D. J. Neven (1991), Cournot competition yields spatial agglomeration, *International Economic Review*, pp. 793–808.

Aramendía, M., C. Larrea, and L. Ruiz (2005), Renegotiation in the repeated Cournot model, *Games and Economic Behavior*, *52*(1), 1–19.

Auerbach, A. J., and Y. Gorodnichenko (2012), Measuring the output responses to fiscal policy, *American Economic Journal: Economic Policy*, *4*(2), 1–27.

Auerbach, A. J., and Y. Gorodnichenko (2014), Fiscal multipliers in japan, *Tech. rep.*, National Bureau of Economic Research.

Baggs, J. (2005), Firm survival and exit in response to trade liberalization, *Canadian Journal of Economics/Revue canadienne d'économique*, *38*(4), 1364–1383.

Bai, H. (2016), Unemployment and credit risk, *Browser Download This Paper*.

Beck, T., A. Demirguc-Kunt, L. Laeven, and R. Levine (2008), Finance, firm size, and growth, *Journal of Money, Credit and Banking*, *40*(7), 1379–1405.

Berger, A. N., A. Demirguc-Kunt, R. Levine, and J. G. Haubrich (2004), Bank concentration and competition: An evolution in the making, *Journal of Money, Credit, and Banking*, *36*(3), 433–451.

Bernanke, B., M. Gertler, and S. Gilchrist (1996), The financial accelerator and the flight to quality, *The Review of Economics and Statistics*, *78*(1), 1–15.

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

Bhuyan, S., and M. McCafferty (2013), Us brewing industry profitability: A simultaneous determination of structure, conduct, and performance, *Journal of agricultural & food industrial organization*, *11*(1), 139–150.

Black, S. E., and P. E. Strahan (2002), Entrepreneurship and bank credit availability, *The Journal of Finance*, *57*(6), 2807–2833.

Bos, J. W., J. W. Kolari, and R. C. Van Lamoen (2013), Competition and innovation: Evidence from financial services, *Journal of Banking & Finance*, *37*(5), 1590–1601.

Breinlich, H. (2008), Trade liberalization and industrial restructuring through mergers and acquisitions, *Journal of international Economics*, *76*(2), 254–266.

Brown, S. (1989), *Retail location theory: The legacy of Harold Hotelling*.

Cabras, I., and C. Bamforth (2016), From reviving tradition to fostering innovation and changing marketing: the evolution of micro-brewing in the uk and us, 1980–2012, *Business History*, *58*(5), 625–646.

Caggiano, G., E. Castelnuovo, and N. Groshenny (2014), Uncertainty shocks and unemployment dynamics in us recessions, *Journal of Monetary Economics*, *67*, 78–92.

Capozza, D. R., and R. Van Order (1982), Product differentiation and the consistency of monopolistic competition: A spatial perspective, *The Journal of Industrial Economics*, pp. 27–39.

Cetorelli, N., and P. E. Strahan (2006), Finance as a barrier to entry: Bank competition and industry structure in local us markets, *The Journal of Finance*, *61*(1), 437–461.

Chang, M.-H., et al. (1991), The effects of product differentiation on collusive pricing, *International Journal of Industrial Organization*, *9*(3), 453–469.

Chava, S., A. Oettl, A. Subramanian, and K. V. Subramanian (2013), Banking deregulation and innovation, *Journal of Financial Economics*, *109*(3), 759–774.

Che, Y.-K., and S.-W. Yoo (2001), Optimal incentives for teams, *The American Economic Review*, *91*(3), 525–541.

Cheng, C. (2016), Moral hazard in teams with subjective evaluations, *Tech. rep.*, Northwestern University.

Coad, A. (2010), The exponential age distribution and the pareto firm size distribution, *Journal of Industry, Competition and Trade*, *10*(3-4), 389–395.

Cockburn, I. M., and R. M. Henderson (2001), Scale and scope in drug development: unpacking the advantages of size in pharmaceutical research, *Journal of health economics*, *20*(6), 1033–1057.

Colantone, I., K. Coucke, and L. Sleuwaegen (2008), Globalization and firm exit: Differences between small and large firms.

Conley, T. G., and E. Ligon (2002), Economic distance and cross-country spillovers, *Journal of Economic Growth*, *7*(2), 157–187.

Cordoba, J.-C., and M. Ripoll (2004), Credit cycles redux, *International Economic Review*, *45*(4), 1011–1046, doi:10.1111/j.0020-6598.2004.00296.x.

Davis, S. J., and J. Haltiwanger (2014), Labor market fluidity and economic performance, *Tech. rep.*, National Bureau of Economic Research.

Decker, R., J. Haltiwanger, R. Jarmin, and J. Miranda (2013), The secular decline in business dynamism in the us, *Manuscript, University of Maryland*.

Decker, R., J. Haltiwanger, R. Jarmin, and J. Miranda (2014), The role of entrepreneurship in us job creation and economic dynamism, *The Journal of Economic Perspectives*, *28*(3), 3–24.

Den Haan, W. J., G. Ramey, and J. Watson (2000), Job destruction and propagation of shocks, *American economic review*, *90*(3), 482–498.

Economides, N. (1996), The economics of networks, *International journal of industrial organization*, *14*(6), 673–699.

Economides, N., et al. (1986), Minimal and maximal product differentiation in hotellings duopoly, *Economics Letters*, *21*(1), 67–71.

Ericson, R., and A. Pakes (1995), Markov-perfect industry dynamics: A framework for empirical work, *The Review of Economic Studies*, *62*(1), 53–82.

Evans, C. L., and J. Harrigan (2005), Distance, time, and specialization: Lean retailing in general equilibrium, *American Economic Review*, *95*(1), 292–313.

Färe, R., S. Grosskopf, B. J. Seldon, and V. J. Tremblay (2004), Advertising efficiency and the choice of media mix: a case of beer, *International Journal of Industrial Organization*, *22*(4), 503–522.

Fuchs, W. (2007), Contracting with repeated moral hazard and private evaluations, *The American Economic Review*, *97*(4), 1432–1448.

Fudenberg, D., and E. Maskin (1986), The folk theorem in repeated games with discounting or with incomplete information, *Econometrica*, *54*(3), 533–554.

George, L. M. (2009), National television and the market for local products: The case of beer, *The Journal of Industrial Economics*, *57*(1), 85–111.

Gilchrist, S., and E. Zakrajšek (2012), Credit spreads and business cycle fluctuations, *American Economic Review*, *102*(4), 1692–1720.

Gokhale, J., and V. J. Tremblay (2012), Competition and price wars in the us brewing industry, *Journal of Wine Economics*, *7*(2), 226–240.

Goldlücke, S., and S. Kranz (2012), Infinitely repeated games with public monitoring and monetary transfers, *Journal of Economic Theory*, *147*(3), 1191–1221.

Goldlücke, S., and S. Kranz (2013), Renegotiation-proof relational contracts, *Games and Economic Behavior*, *80*, 157–178.

Goldmanis, M., A. Hortaçsu, C. Syverson, and Ö. Emre (2010), E-commerce and the market structure of retail industries, *The Economic Journal*, *120*(545), 651–682.

Gu, W., G. Sawchuk, and L. W. Rennison (2003), The effect of tariff reductions on firm size and firm turnover in canadian manufacturing, *Review of World Economics*, *139*(3), 440–459.

Guillén, J. B., and S. Pinto (2007), Bank branching deregulation: a spatial competition model, in *Annals of economical and business studies*, 17, pp. 87–108, Publications Service.

Hagedorn, M., and I. Manovskii (2008), The cyclical behavior of equilibrium unemployment and vacancies revisited, *American Economic Review*, *98*(4), 1692–1706.

Hairault, J.-O., F. Langot, and S. Osotimehin (2010), Matching frictions, unemployment dynamics and the cost of business cycles, *Review of Economic Dynamics*, *13*(4), 759 – 779, doi:http://dx.doi.org/10.1016/j.red.2010.05.001.

Hanson, G. H. (2005), Market potential, increasing returns and geographic concentration, *Journal of international economics*, *67*(1), 1–24.

Harrington, J. E., and A. Skrzypacz (2007), Collusion under monitoring of sales, *The RAND Journal of Economics*, *38*(2), 314–331.

Harrington, J. E., and A. Skrzypacz (2011), Private monitoring and communication in cartels: Explaining recent collusive practices, *The American Economic Review*, *101*(6), 2425–2449.

Hart, O., and B. Holmström (1986), *The Theory of Contracts*, Department of Economics, Massachusetts Institute of Technology.

Head, K., and J. Ries (1999), Rationalization effects of tariff reductions, *Journal of International Economics*, *47*(2), 295–320.

Herman, E. (2001), Independent new york city bookstore reaches final chapter, *New York Daily News*.

Holmström, B. (1982), Moral hazard in teams, *The Bell Journal of Economics*, *13*(2), 324–340.

Hooper, W., and M. K. Rawls (2014), Borders group, inc.s final chapter: How a bookstore giant failed in the digital age.

Hopenhayn, H., and R. Rogerson (1993), Job turnover and policy evaluation: A general equilibrium analysis, *Journal of political Economy*, pp. 915–938.

Hotelling, H. (1929), Stability in competition, *The Economic Journal*, *39*(153), 41–57.

House, C. L. (2006), Adverse selection and the financial accelerator, *Journal of Monetary Economics*, *53*(6), 1117–1134.

Huggett, M. (1993), The risk-free rate in heterogeneous-agent incomplete-insurance economies, *Journal of economic Dynamics and Control*, *17*(5), 953–969.

Hurwicz, L. (2008), But who will guard the guardians?, *The American Economic Review*, *98*(3), 577–585.

Irmen, A., J.-F. Thisse, et al. (1998), Competition in multi-characteristics spaces: Hotelling was almost right, *Journal of economic theory*, *78*(1), 76–102.

Jermann, U., and V. Quadrini (2012), Macroeconomic effects of financial shocks, *The American Economic Review*, *102*(1), 238–271.

Jordà, Ò. (2005), Estimation and inference of impulse responses by local projections, *American economic review*, pp. 161–182.

Kiyotaki, N., and J. Moore (1997), Credit cycles, *Journal of Political Economy*, *105*(2), 211–48.

Kroszner, R. S., and P. E. Strahan (1999), What drives deregulation? economics and politics of the relaxation of bank branching restrictions, *The Quarterly Journal of Economics*, *114*(4), 1437–1467.

Krugman, P. (1991), Increasing returns and economic geography, *Journal of political economy*, *99*(3), 483–499.

Krusell, P., and A. A. Smith (1999), On the welfare effects of eliminating business cycles, *Review of Economic Dynamics*, *2*(1), 245–272.

Krusell, P., and A. A. Smith, Jr (1998), Income and wealth heterogeneity in the macroeconomy, *Journal of Political Economy*, *106*(5), 867–896.

Kvasnička, M., R. Staněk, and O. Krčál (2018), Is the retail gasoline market local or national?, *Journal of Industry, Competition and Trade*, *18*(1), 47–58.

Lagos, R., and G. Rocheteau (2009), Liquidity in asset markets with search frictions, *Econometrica*, *77*(2), pp. 403–426.

Lambrecht, A.-U., et al. (2006), Adoption and usage of online services in the presence of complementary offline services: retail banking.

Lancaster, K. (1982), Innovative entry: Profit hidden beneath the zero, *The Journal of Industrial Economics*, pp. 41–56.

Lee, B., and V. J. Tremblay (1992), Advertising and the us market demand for beer, *Applied Economics*, *24*(1), 69–76.

Levine, R. (2005), Finance and growth: theory and evidence, *Handbook of economic growth*, *1*, 865–934.

Lileeva, A. (2008), Trade liberalization and productivity dynamics: evidence from canada, *Canadian Journal of Economics/Revue canadienne d'économique*, *41*(2), 360–390.

Liu, Z., P. Wang, and T. Zha (2013), Land-price dynamics and macroeconomic fluctuations, *Econometrica*, *81*(3), 1147–1184, doi:10.3982/ECTA8994.

Long, N. V., and N. Vousden (1995), The effects of trade liberalization on cost-reducing horizontal mergers, *Review of International Economics*, *3*(2), 141–155.

Lucas, R. E. (1978), On the size distribution of business firms, *The Bell Journal of Economics*, pp. 508–523.

Luttmer, E. G. (2007), Selection, growth, and the size distribution of firms, *The Quarterly Journal of Economics*, *122*(3), 1103–1144.

Mailath, G. J., V. Nocke, and L. White (2017), When and how the punishment must fit the crime, *International Economic Review*, *58*(2), 315–330.

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

Melitz, M. J., and G. I. Ottaviano (2008), Market size, trade, and productivity, *The review of economic studies*, *75*(1), 295–316.

Molloy, R., R. Trezzi, C. L. Smith, and A. Wozniak (2016), Understanding declining fluidity in the us labor market, *Brookings Papers on Economic Activity*, *2016*(1), 183–259.

Mueller, W. F., and L. G. Hamm (1974), Trends in industrial market concentration, 1947 to 1970, *The Review of Economics and Statistics*, pp. 511–520.

Mukhopadhyay, A. K. (1985), Technological progress and change in market concentration in the us, 1963-77, *Southern Economic Journal*, pp. 141–149.

Neftci, S. N. (1984), Are Economic Time Series Asymmetric over the Business Cycle?, *Journal of Political Economy*, *92*(2), 307–28.

Norman, G., and N. K. Nichols (1982), Dynamic market strategy under threat of competitive entry: an analysis of the pricing and production policies open to the multinational company, *The Journal of Industrial Economics*, pp. 153–174.

Ostrom, E., J. Walker, and R. Gardner (1992), Covenants with and without a sword: Self-governance is possible., *American Political Science Review*, *86*(2), 404–417.

Pagano, P., and F. Schivardi (2003), Firm size distribution and growth, *The Scandinavian Journal of Economics*, *105*(2), 255–274.

Pakes, A., and P. McGuire (2001), Stochastic algorithms, symmetric markov perfect equilibrium, and the curseof dimensionality, *Econometrica*, *69*(5), 1261–1281.

Petersen, M. A., and R. G. Rajan (2002), Does distance still matter? the information revolution in small business lending, *The journal of Finance*, *57*(6), 2533–2570.

Petrosky-Nadeau, N. (2013), Tfp during a credit crunch, *Journal of Economic Theory*, *148*(3), 1150–1178.

Petrosky-Nadeau, N., and E. Wasmer (2012), The cyclical volatility of labor markets under frictional financial markets, *Available at SSRN 1553108*.

Petrosky-Nadeau, N., and L. Zhang (2013a), Unemployment crises, *Working Paper 19207*, National Bureau of Economic Research.

Petrosky-Nadeau, N., and L. Zhang (2013b), Unemployment crises, *Tech. rep.*, National Bureau of Economic Research.

Phlips, L., and J.-F. Thisse (1982), Spatial competition and the theory of differentiated markets: an introduction, *The Journal of Industrial Economics*, *31*(1/2), 1–9.

Pugsley, B. W., and A. Sahin (2015), Grown-up business cycles, *US Census Bureau Center for Economic Studies Paper No. CES-WP-15-33*.

Rahman, D. (2012), But who will monitor the monitor?, *The American Economic Review*, *102*(6), 2767–2797.

Reedy, E., and R. J. Strom (2012), Starting smaller; staying smaller: Americas slow leak in job creation, in *Small Businesses in the Aftermath of the Crisis*, pp. 71–85, Springer.

Rollinger, M. D. (1996), Interstate banking and branching under the riegle-neal act of 1994, *Harv. J. on Legis.*, *33*, 183.

Rouwenhorst, K. G. (1995), Asset pricing implications of equilibrium business cycle models, in *Frontiers of business cycle research*, edited by T. F. Cooley, Princeton University Press, Princeton.

Senzel, J. (1992), The mcfadden act, *Interstate Branch Banking Reform: Preserving the Policies Underlying, Boston University Law Review*.

Sinai, T., and J. Waldfogel (2004), Geography and the internet: Is the internet a substitute or a complement for cities?, *Journal of Urban Economics*, *56*(1), 1–24.

Stahl, K. (1982), Differentiated products, consumer search, and locational oligopoly, *The Journal of Industrial Economics*, pp. 97–113.

Strahan, P. (2003), The real effects of us banking deregulation, *Federal Reserve Bank of St. Louis Review*, *85*(4), 111.

Swaminathan, A. (1998), Entry into new market segments in mature industries: Endogenous and exogenous segmentation in the us brewing industry, *Strategic Management Journal*, pp. 389–404.

Symeonidis, G. (1996), Innovation, firm size and market structure.

Tremblay, V. J., N. Iwasaki, and C. H. Tremblay (2005), The dynamics of industry concentration for us micro and macro brewers, *Review of Industrial Organization*, *26*(3), 307–324.

Tsang, E. W., and P. S. Yip (2007), Economic distance and the survival of foreign direct investments, *Academy of Management Journal*, *50*(5), 1156–1168.

Wasmer, E., and P. Weil (2004), The macroeconomics of labor and credit market imperfections, *American Economic Review*, *94*(4), 944–963.

Wasmer, E., and Y. Zenou (2006), Equilibrium search unemployment with explicit spatial frictions, *Labour Economics*, *13*(2), 143–165.

Weintraub, G. Y., C. L. Benkard, and B. Van Roy (2008), Markov perfect industry dynamics with many firms, *Econometrica*, *76*(6), 1375–1411.

Weisman, R. (2004), Final chapter for wordsworth books on bostons harvard square is saturday, *Boston Globe*.

Yitzhaki, S. (1994), Economic distance and overlapping of distributions, *Journal of Econometrics*, *61*(1), 147–159.

Zentner, A. (2008), Online sales, internet use, file sharing, and the decline of retail music specialty stores, *Information Economics and Policy*, *20*(3), 288–300.

Zhu, K. (2001), Internet-based distribution of digital videos: the economic impacts of digitization on the motion picture industry, *Electronic Markets*, *11*(4), 273–280.