Markup Cycles, Dynamic Misallocation, and Amplification

Marcus M. Opp*[a], Christine A. Parlour[a], and Johan Walden[a]

[a]University of California at Berkeley, Haas School of Business, USA

July 9, 2014

Abstract

We develop a tractable dynamic general equilibrium model of oligopolistic competition with a continuum of heterogeneous industries. Industries are exposed to aggregate and industry-specific productivity shocks. Firms in each industry set value-maximizing state-contingent markups, taking as given the behavior of all other industries. When consumers are risk-averse, industry markups are countercyclical with regards to the industry-specific component, but may be procyclical with regards to the aggregate shock. The general equilibrium dispersion of markups implied by the optimization of heterogeneous industries creates misallocation of labor across industries. The misallocation, in turn, generates aggregate welfare losses state-by-state that feed back into the industry problem via a representative agent's marginal utility of aggregate consumption. Misallocation dynamics may transmit industry-specific shocks, or amplify small aggregate shocks, and so lead to aggregate fluctuations through these feedback effects.

JEL classification: L16, E32, L13

Keywords: Oligopolistic Competition, Markup Cycles, Allocative Efficiency, Dynamic Games, DSGE Models.

*Corresponding author. E-mail: mopp@haas.berkeley.edu, 545 Student Services Building #1900, CA 94720-1900. *Phone:* +1-510-643-0658. *Fax:* +1-510-643-1420.

1 Introduction

How does industry-level firm strategic interaction influence the aggregate economy? Although the effects of strategic interaction have been thoroughly analyzed in the Industrial Organization literature, the aggregate implications have typically been ignored. In this paper we develop a general equilibrium model in which *oligopolistic* intra-industry competition generates markup dispersion across *heterogeneous* industries, which leads to resource misallocation (see Lerner [34]) and hence affects aggregate consumption. Following standard asset pricing insights, changes in aggregate consumption affect agents' marginal utilities across states and thereby the valuation of firms' future cash flows; this in turn feeds back into the firms' ability to sustain collusion, leading to a rich set of implications.

We study a discrete time, infinite horizon general equilibrium economy with a continuum of industries, each of which is defined by a production technology. Within each industry, a finite number of identical strategic firms hire labor to produce a homogeneous good. The price of the good in each industry is determined by the outcome of a dynamic pricing game similar to Rotemberg and Saloner [43]. A representative agent consumes all goods, supplies all labor, and owns all the firms; thus all profits are valued by her preferences over consumption. We allow industries to differ cross-sectionally, both in their number of firms and their exposure to productivity shocks. These sources of heterogeneity allow us to capture industry-specific strategic behavior, generate heterogeneous markups, and analyze how industry-specific productivity shocks are transmitted to the aggregate economy.

Firms in each industry maximize profits subject to intertemporal incentive compatibility constraints: In each period, each firm weighs the value of high short-term profits that can be obtained by aggressive pricing against the long-term profits that are obtained when all firms cooperate. The value of such long-term profits is determined by the preferences of the representative agent. In general equilibrium, the representative agent's consumption bundle depends on the sum of all outputs produced in each industry. If markups are heterogeneous across industries, relative goods' prices are distorted compared to the first-best outcome, leading to a) misallocation of labor to industries and b) a reduction in aggregate consumption. Such changes in consumption affect the representative agent's marginal utility across states and hence her valuation of each industry's profits, and therefore feed back into each firm's ability to sustain collusion. Thus, while each industry takes the macro dynamics as given, industries jointly affect these macro dy-

namics through changes in the representative agent's consumption. Our paper therefore provides a tight link between strategic industry behavior and aggregate outcomes.

We make three theoretical contributions. First, we focus on one industry. We characterize markups and derive conditions under which they are procyclical and countercyclical, respectively. Countercyclical markups are often associated with oligopolistic competition, based on Rotemberg and Saloner [43]. In their framework, high product demand in good times increases firms' incentives to undercut competitors to reap immediate rewards; therefore equilibrium markups narrow in good times. Our paper shows that this intuition can be overturned. Our arguments follow from the fundamental insights of consumption based asset pricing that market discount rates vary with the state of the economy, in contrast to the risk-neutral setting of Rotemberg and Saloner [43]. If discount rates are sufficiently low in good times, then the present value of future cooperation compared to current period profits is higher in booms, making procyclical markups possible. This insight is general. Within our model, market discount rates can be endogenously countercyclical if the representative agent's intertemporal elasticity of substitution is low. With constant relative risk aversion, the threshold level for procyclicality is given by a coefficient of relative risk aversion of 1, i.e., for logarithmic utility.

While the cyclicality of the "average industry" is ambiguous, following the previous logic, we also show that one can decompose an industry's profit variations into an aggregate and an industry-specific component and that the source of ambiguity lies in the aggregate component. Markups are always countercyclical with respect to the industry-specific component, i.e., controlling for the aggregate shock. This is natural, since industry-specific shocks do not affect the marginal utility of consumption and hence discount rates.

It is important to understand how and why markups vary over the business cycle in the design of optimal monetary policy. The cyclicality of markups is a key building block of leading Neo-Keynesian macroeconomic models (see e.g., Goodfriend and King [25], Woodford [49], and Christiano et al. [13]). As Nekarda and Ramey [38] highlight, most Neo-Keynesian models share the feature that markups fall in response to positive demand shifts, while providing empirical evidence that this prediction does not hold up in the US post-war data: Average markups are slightly procyclical (see Figure 1 of Nekarda and Ramey [38]). Using asset-pricing insights, our model can generate procyclical markups for the average industry using reasonable parameter values. In addition, our multi-industry framework allows for the possibility of heterogeneous markup cyclicality across industries. We provide first-pass evidence that this heterogeneity is empirically relevant. We estimate

a panel of price-cost margins (PCM) for 451 industries between 1959 and 2009 using the NBER manufacturing productivity database of Bartelsman and Gray.[1] Figure 1 plots the resulting histogram of time-series correlation coefficients of industry markups with industrial output growth (GDP). Some industries exhibit strong countercyclical markups while others exhibit strong procyclical markups, a pattern that our model can replicate by allowing for industry-specific shocks.[2]

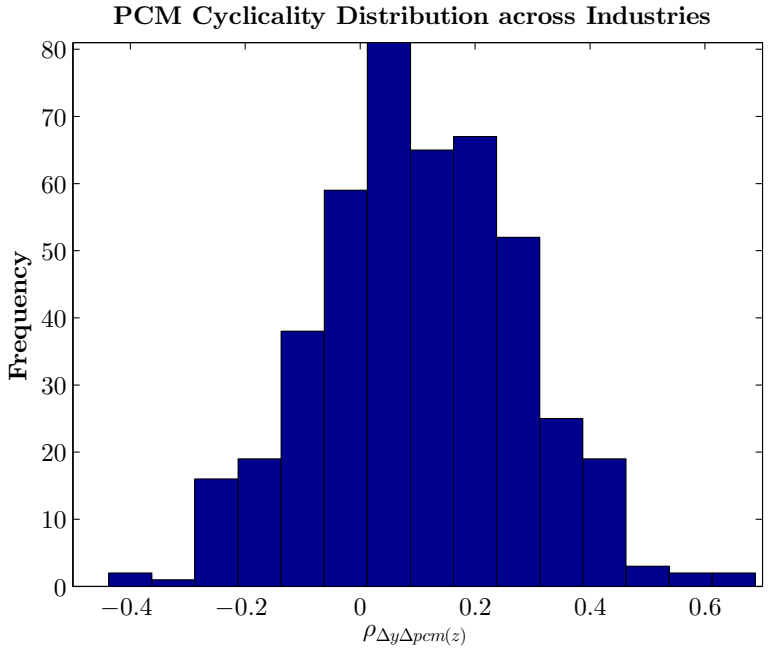

Figure 1. This graph plots a histogram of the distribution of markup cyclicality across industries. Specifically, the term $\rho_{\Delta y \Delta pcm(z)}$ refers to the time-series correlation coefficient of yearly log changes of the price cost margin of a particular industry z with yearly log changes in industrial output. Since the average industry features $\rho_{\Delta y \Delta pcm(z)} > 0$, the evidence suggests slightly procyclical markups. See Appendix A for data description and variable definitions.

Our second, theoretical contribution is to analyze how the heterogeneous oligopolistic industry-level firm behavior may amplify technological shocks or even be the only source of aggregate volatility in the economy. *Misallocation* arises because incentive constraints of heterogeneous industries are not synchronized across *industries*, either due to industry-specific shocks or different levels of competitiveness. *Misallocation dynamics* occur because the heterogeneity of the incentive problem (and hence markups) varies

[1]While (average) price cost margins only correspond to precise markup estimates under special assumptions, e.g., if labor is the only factor input and production is constant returns to scale, they should be interpreted as a reasonable first pass proxy. (See Nekarda and Ramey [38] for more advanced methods.)

[2]See also Bils et al. [8], who provide evidence on variation of relative markups of durables and nondurables over the business cycle.

across *states*. While misallocations originate in industry-specific shocks, interesting feedback effects may arise. Small changes in a few industries may become amplified if they affect other industries' ability to sustain collusive outcomes through the effects they have on the representative agents future valuation of consumption. In several examples we show that the amplification effects can be large. We also highlight that shock amplification occurs whenever the endogenous cross-sectional dispersion of markups is higher during recessions than in good times, and that dampening of shocks is also theoretically possible in equilibrium, if markup dispersion is sufficiently procyclical.

Our third contribution is technical: We characterize the existence and qualitative behavior of equilibrium in our model. Given the complete generality of our set-up, allowing for full heterogeneity across industries and states, existence of equilibrium is by no means clear, a priori. Our main result in this part of the paper is Proposition 4, which shows the existence of equilibrium under minimal assumptions.

Literature We are certainly not the first researchers to address these issues and to explore micro foundations of macro shocks. Furthermore, as our approach straddles multiple fields, it draws on various literatures including the industrial organization literature, the literature on misallocations and the literature on the propagation of macro shocks.

Our partial equilibrium results are most closely related to the Industrial Organization literature on strategic competition over the business cycle following the seminal paper by Rotemberg and Saloner [43] (see, e.g., Chevalier and Scharfstein [11], Chevalier and Scharfstein [12], Bagwell and Staiger [4], and Haltiwanger and Harrington [26]). Synthesizing the literature and our contribution in a nutshell, one can identify three distinct and intuitive channels governing the cyclicality of markups: current period industry demand, future industry demand growth, and discount rates. Rotemberg and Saloner [43] find that higher *current period* demand (ceteris paribus) increases the incentive to deviate and lowers equilibrium markups. Haltiwanger and Harrington [26] as well as Bagwell and Staiger [4] make the important observation that higher *future demand* (growth) decreases the incentive to deviate since high future collusion profits make deviation today less attractive. The discount rate channel that we propose affects the tradeoff between today's profits and future profits: In good aggregate states, the representative agent values an additional consumption unit less than in bad times, which effectively lowers her discount rate and hence increases valuations, systematically leading to (more) procyclical markups for the average industry.[3]

[3]dal Bo [14] considers stochastic interest rates in a collusion model, but since these fluctuations are

We extend this partial equilibrium literature by incorporating strategic behavior into a general equilibrium framework with multiple industries, thereby endogenizing aggregate consumption and the pricing of risk. Our general equilibrium framework is built on the seminal paper by Rotemberg and Woodford [44] although the analysis focuses on different effects of markups. In their model, countercyclical markups can transmit aggregate demand shocks by the government to the real economy. In our paper, we shut down the real effects of markup *levels* by excluding government expenditures and assuming inelastic labor supply. Instead, we focus on the misallocation resulting from markup *dispersion*, which is absent from their model due to the assumption of symmetric industries. Our extension to allow for cross sectional variation of industry concentration and productivity makes it possible to generate dynamics of markup *dispersion* in a completely real model, microfounded by value-maximizing strategic behavior at the industry level.

Since misallocations are the only source of inefficiencies in our general equilibrium framework, our paper features similar distortions as classical sticky-price models in the spirit of Calvo [9]. In contrast to sticky-price models, however, prices in our model are fully flexible and are determined *endogenously* as the outcome of a strategic game of optimizing, heterogeneous industries. As Bilbiie et al. [6] point out, the fundamental economics behind misallocation can be traced back to early essays of Lerner [34] and Samuelson [45]. Misallocation of labor via markup dispersion is particularly relevant for the literature on international trade since competition from abroad naturally affects industries in a heterogeneous way (see Epifani and Gancia [20], Holmes et al. [27], Edmond et al. [19], and Dhingra and Morrow [16]). From a modeling perspective, the literature on misallocation also highlights the special role of CES preferences under monopolistic competition in that market outcomes are efficient due to markups synchronization (see in particular Bilbiie et al. [6] and Dhingra and Morrow [16]).[4] Instead, our paper shows that inefficiencies can arise even in settings with CES preferences (and inelastic labor supply) by allowing for oligopolistic competition with heterogeneous industries. This allows us to keep the tractability and standard aggregation results of CES preferences, while being able to match relevant heterogeneity across industries.

Empirical studies suggest that losses from misallocation can be quantitatively large; at least in emerging market countries. Hsieh and Klenow [28] estimate static losses ranging from $30\% - 50\%$ in China and $40\% - 60\%$ in India. In a dynamic setting, Peters [41] considers the joint effect of misallocation, endogenous entry (see also Bilbiie et al. [7]) and

exogenous (i.i.d.), the paper does not address pro- or countercyclicality of markups.

[4]See Zhelobodko et al. [51] for a generalization of CES preferences.

incentives to innovate (see also Kung and Schmid [32]). Using a sample of manufacturing firms in Indonesia, he finds that a large proportion of the welfare gains from reducing barriers to entry results from the effect on the equilibrium growth rate rather than the reduction in (static) misallocation.

Since our paper combines real technology shocks with the just described endogenous misallocations, our paper also relates to an extensive literature on business cycles (e.g., Kydland and Prescott [33]; Long and Plosser [35]; Gabaix [22]; Acemoglu et al. [2]). In contrast to the real business cycle literature, however, significant aggregate fluctuations may arise even when aggregate "technological" shocks are small. A recent strand of literature has aimed at explaining how technological shocks at the individual firm or industry level do not diversify out, but may affect aggregate productivity. Gabaix [22] notes that if the distribution of firm size is heavy-tailed, firm-specific shocks may indeed affect aggregate productivity. Acemoglu et al. [2], suggest that inter-sectoral input-output linkages between industries may lead to "cascades effects" where a shock in one industry spreads through the economy and thereby becomes an aggregate shock. In our setup, such "cascade effects" may arise through the channel of the pricing kernel even if there is no direct input-output linkage between sectors. The mechanism in our model is also quite different, more along the lines suggested in Jovanovic [31], who shows that idiosyncratic shocks may not cancel out in strategic games with a large number of players. We develop examples in which aggregate productivity is close to constant across states, but because it varies at the sectoral level, the strategic behavior of firms leads to aggregate shocks in equilibrium.

Our results highlight how strategic interaction between firms can generate endogenous fluctuations. These results are related to Gali [24] and Schmitt-Grohe [46] who, building on Woodford [48] and Woodford [50], study stationary sunspot equilibria in models with markups and investments. Both papers focus on the symmetric case with monopolistic competition, in which case the multiplicity of equilibria arises because of self-fulfilling expectations about future growth rates.[5] In contrast, our model features a unique equilibrium under symmetric behavior, i.e., homogeneous industries. Our key contribution is to allow for heterogeneous sectors in which welfare distortions arise from the dispersion of markups across industries. Multiplicity of equilibria can only occur if feedback effects are sufficiently strong.

The rest of the paper is organized as follows. In Section 2 we present the economic

[5]In Jaimovich [29], sunspot equilibria and countercyclical markups arise via entry and exit decisions (also see Jaimovich and Floetotto [30]).

framework of the model. The equilibrium analysis of each industry and their joint effect on aggregate outcomes is presented in Section 3. Section 4 shows the existence of general equilibrium under general conditions and discusses how endogenous misallocation dynamics may arise. Section 5 discusses the empirical implications of our paper. All proofs are delegated to the Appendix.

2 Model Framework

2.1 Physical Environment

Consider an infinite horizon, discrete time, discrete state economy in which time is indexed by $t \in \mathbb{Z}_+$ and the time t state of the world is denoted by $s_t \in \{1, 2, \dots S\}$.[6] Each period there is a transition between states, which is governed by a Markov process with time invariant transition probabilities:

$$\mathbb{P}(s_{t+1} = j | s_t = i) = \Phi_{i,j}. \tag{1}$$

Here, $\Phi_{i,j}$ refers to the element on the ith row and jth column of the matrix $\Phi \in \mathbb{R}_+^{S \times S}$. We assume that Φ is irreducible and aperiodic, so that the process has a unique long-term stationary distribution.

2.1.1 Production

There is a continuum of industries, indexed by $z \in [0, 1]$, each consisting of $N(z) \geq 1$ identical strategic firms that produce and sell a unique non-storable consumption good. The nature of the strategic environment is discussed in Section 2.2. The production technology for each good z at time t is linear in labor with stochastic productivity $A(z, t) = A_{s_t}(z)(1 + g)^t$. Here, with some abuse of notation, $A_{s_t}(z)$ represents a state-dependent and sector-specific productivity component, whereas $g \geq 0$ represents a common long-term productivity growth rate across all sectors. For ease of exposition, we set $g = 0$ in the main text and refer the reader to Appendix C, which shows the minor modifications necessary for the general case $g > 0$. Also, for tractability we assume that $A : S \times [0, 1] \to \mathbb{R}_{++}$ is a function that satisfies standard integrability conditions so that

[6] Here, $\mathbb{Z}_+ = \{0\} \cup \mathbb{N} = \{0, 1, \dots\}$ is the set of non-negative integers. Also, we follow the standard convention that $\mathbb{R}_+$ is the set of nonnegative real numbers, whereas $\mathbb{R}_{++}$ is the set of strictly positive real numbers.

aggregation across industries is possible. Labor is supplied inelastically by a representative agent, who in each period allocates her one unit of human capital across industries, earning a competitive wage, $w(t)$, in return.[7]

2.1.2 Preferences / Demand

The representative agent possesses iso-elastic preferences over aggregate consumption with risk aversion parameter γ and subjective discount factor δ, i.e.,

$$U = \mathbb{E}\left[\sum_{t=0}^{\infty} \delta^t \frac{C(t)^{1-\gamma}}{1-\gamma}\right], \tag{2}$$

where $C(t)$ represents the Dixit-Stiglitz *CES* consumption aggregator of goods (see Dixit and Stiglitz [17]).[8] Thus,

$$C(t) = \left(\int_0^1 c(z,t)^{\frac{\theta-1}{\theta}} dz\right)^{\frac{\theta}{\theta-1}}. \tag{3}$$

The parameter $\theta > 1$ is the (constant) elasticity of substitution across goods. While industries are thus assumed to be symmetric on the demand side, a more general state dependent utility specification can be easily mapped into our model, which would allow us to capture industry heterogeneity in demand, say cyclical vs. countercyclical goods.[9]

The *CES* specification leads to standard *period-by-period* demand functions as a function of prices $p(z,t)$ and real income $y(t)$:[10]

$$c(z,t) = \frac{y(t)}{p(z,t)^{\theta} P(t)^{-\theta}}, \tag{4}$$

[7]We deliberately shut down the channel of endogenous labor supply to sharpen our findings of factor misallocation across heterogeneous sectors. Thus, our production factor in fixed supply could also be interpreted as "land" that has to be allocated to different sorts of crops (industries). We excluded physical capital accumulation from our model to avoid the issue of disentangling effects of dynamic investment decisions from the effects of state-contingent markups.

[8]See van Binsbergen [47] or Ravn et al. [42] for using *CES* preferences in a dynamic context.

[9]Consider the more general $\tilde{C}(t) = \left(\int_0^1 v_{s_t}(z) c(z,t)^{\frac{\theta-1}{\theta}} dz\right)^{\frac{\theta}{\theta-1}}$ as in Opp [39]. The state dependent "taste" function $v_s(z)$ can then easily be reduced to the case where $v_s(z) \equiv 1$, by transforming the productivity, $A_s(z) \mapsto v_s(z)^{(\theta-1)/\theta} A_s(z)$. Such a transformation can be interpreted as a numeraire change, where the amount of a unit of goods is redefined in each state.

[10]The demand functions $c(z,t)$ yield maximal $C(t)$ given an arbitrary price vector $p(z,t)$ and income $y(t)$. They are obtained via simple first-order conditions.

where $P(t) \equiv \left(\int_0^1 p(z,t)^{1-\theta} dz\right)^{\frac{1}{1-\theta}}$ can be interpreted as the aggregate price index. Without loss of generality, we can normalize the nominal price index $P(t)$ to 1. Hence, all variables are measured in units of aggregate consumption. In particular, real income, $y(t)$, is derived from wages, and distribution of firm profits, $\pi(z,t)$, across all sectors z:

$$y(t) = w(t) + \int_0^1 \pi(z,t) dz, \tag{5}$$

$$\pi(z,t) = \left[p(z,t) - \frac{w(t)}{A(z,t)}\right] c(z,t). \tag{6}$$

2.2 Strategic Environment

Within each industry z, $N(z)$ identical firms play a dynamic Bertrand pricing game with perfect public information, taking as given the behavior of all other industries. In contrast to Rotemberg and Saloner [43], we assume that firm value is determined by the preferences of a risk-averse (rather than risk-neutral) representative agent.

The timing of the stage game in each period, t, is as follows. First, the state, s_t is revealed. Then all firms $i \in \{1, 2, \ldots N(z)\}$ in industry z simultaneously announce their gross markup, $M^{(i)}(z,t)$. For tractability, we express each firm's strategy in terms of gross markups instead of prices, satisfying $p^{(i)}(z,t) = M^{(i)}(z,t)\frac{w(t)}{A_{s_t}(z)}$. Consumers demand the product from the producer with the lowest markup. If all firms announce the same M, total demand in sector z is evenly shared between all $N(z)$ firms. The firms then hire workers at a competitive wage $w(t)$ to meet demand.

Following Abreu [1], we are interested in industry equilibria that generate the highest present value of *industry* profits sustainable by credible threats. We restrict attention to *symmetric*, *pure strategy subgame perfect* equilibria. Firms condition their action at time t on the entire history of past actions of industry z and states up to time t. The relevant history of each industry z, h_t is defined as the entire sequence of markups, states, and aggregate variables:

$$h_t = \left\{\left\{M^{(i)}(z,\tau)\right\}_{i=1}^{N(z)}, s_\tau, P(\tau), y(\tau)\right\}_{\tau=0}^{t}, \tag{7}$$

with h_0 representing the empty history. Thus, a time-t, industry-z strategy for firm i is a mapping from $h_{t-1} \times S$ to a chosen markup, $M^i(z,\tau)$, $f_t^i : h_{t-1} \times S \to R_{++}$, (i.e., $f_t^i \in R_{++}^{h_{t-1} \times S}$). Here, the second parameter, $s \in S$, represents time t information about the state, which is available for the firm. A strategy for firm i is a sequence of time τ

strategies, $\{f_\tau^i\}_{\tau=0}^{\infty}$.

The entire set of subgame perfect equilibria can be enforced with the threat of the worst possible subgame perfect equilibrium. In our environment, the most severe punishment is given by the perfectly competitive outcome, i.e., zero profits forever after a deviation. Therefore, any subgame perfect equilibrium must satisfy the following incentive constraints at each date t,

$$\frac{\pi_t(z) + V_t(z)}{N(z)} \geq \pi_t(z). \tag{8}$$

That is, collusion is only sustainable if each firm's share, $\frac{1}{N(z)}$, of today's industry profits, $\pi_t(z)$, and the present value of future industry profits, V_t, is greater or equal to the best-possible one period deviation of capturing the entire industry demand π_t and zero profits thereafter. An important force of this incentive constraint in our setup is captured by the valuation of uncertain profit streams by a risk-averse agent which will be reflected in $V_t(z)$ (see detailed discussion in Section 3.2.2).

Myopic industry value maximization of $V_t(z)$, subject to equation 8, represents the only friction in our economy.[11] While the equilibrium outcome of this game is in general non-trivial (see Section 3.3), the two polar cases of a monopoly, i.e., $N(z) = 1$, and perfect competition provide useful bounds. If the industry is served by a monopolist, he maximizes industry profits (equation 6) subject to consumer demand (equation 4) which leads to an optimal markup of:

$$M^m(z,t) = M^m = \frac{\theta}{\theta - 1}. \tag{9}$$

If, on the other hand, $N(z)$ is infinite, then we expect prices to be set competitively. In this case, the markup is 1. If the number of firms is finite but greater than one, we expect equilibrium markups to be somewhere in between the competitive and monopolistic prices, i.e., $M \in \left[1, \frac{\theta}{\theta-1}\right]$.

[11] We are implicitly assuming that firms can coordinate within an industry to achieve this best outcome with this equilibrium selection mechanism. This trivially rules out any outcomes where markups are higher than $\frac{\theta}{\theta-1}$, and outcomes where markups are lower than necessary. We do *not*, however, assume that firms can coordinate across industries, since in a large economy there are many industries and global coordination therefore is typically not possible.

3 Partial Equilibrium Analysis

Our partial equilibrium analysis consists of two parts. First, for an arbitrary exogenous distribution of markups across industries, we characterize aggregate consumption, and show that it, together with a measure of aggregate markups, determines the efficiency losses in the economy (Section 3.2). Second, given the aggregate consumption and aggregate markup dynamics, we solve for the partial equilibrium outcome of one sector z in the economy, i.e., the optimal state-contingent markups (Section 3.3).

3.1 Preliminaries

We focus on equilibria which are time invariant in that equilibrium outcomes are the same at t_1 and t_2 if the states are the same, i.e., if $s_{t_1} = s_{t_2}$. Hence, we introduce the following notation for equilibrium markups (and similarly for other variables):

$$M(z,t) = M_{s_t}(z). \tag{10}$$

The focus on time invariant equilibria is natural in our stationary environment, since we prove that optimizing firm behavior in one particular industry is endogenously time invariant provided that all other industries exhibit time-invariant behavior. Moreover, it is ensured that (at least) one time-invariant equilibrium exists (see Proposition 4). We want to emphasize that this formulation does not impose any restriction on *off-equilibrium path* behavior.

For ease of exposition, we decompose productivity shocks $A_s(z)$ into the functions $\alpha_s(z)$ and $\bar{A}_s$ where $\alpha : S \times [0,1]$ and the vector $\bar{A} \in \mathbb{R}^S_+$. Specifically,

$$\alpha_s(z) \equiv \frac{A_s(z)^{\theta-1}}{\int_0^1 A_s(z)^{\theta-1} dz} = \left(\frac{A_s(z)}{\bar{A}_s}\right)^{\theta-1}, \qquad \text{where} \tag{11}$$

$$\bar{A}_s \equiv \left[\int_0^1 A_s(z)^{\theta-1} dz\right]^{\frac{1}{\theta-1}}. \tag{12}$$

Here, $\bar{A}$ represents the average productivity shock to the economy and $\alpha_s(z)$ captures the industry productivity shock relative to the economy. In other words, changes in $\alpha(z)$ across states are *industry-specific* shocks, whereas changes in $\bar{A}$ are *aggregate* shocks. We can also view $\alpha(z)$ as an S-vector, $\alpha(z) \in \mathbb{R}^S$. Note that an industry with a constant α across all states, moves one-to-one with the aggregate state. Since industries are of

infinitesimal size, industry-specific shocks α can thus also be interpreted as idiosyncratic shocks. As a result of the normalization, the average relative industry state is equal to one, i.e., $\int_0^1 \alpha_s(z)dz = 1$. Now instead of specifying A, we can equivalently specify the function of industry-specific shocks, α, and the vector of aggregate shocks, $\bar{A} \in \mathbb{R}^S_{++}$. Given the previous argument, the exogenous variables in the economy can then be represented by the tuple $\mathcal{E} = (\alpha, \bar{A}, N, \Phi, \theta, \gamma, \delta)$.

3.2 Aggregate Consumption and Welfare

Aggregate consumption is an important endogenous variable. As outlined above, we will first treat the outcome of the strategic game for each industry and each state as exogenously given, as summarized by the gross markup functions for each industry, $M_s(z)$. Together with the exogenous functions, $\alpha_s(z)$ and $\bar{A}_s$, the real outcome in the economy or the consumer's consumption bundle is completely determined, state-by-state. We will use aggregate consumption in two ways. First, as a measure of welfare and, second, to value a stream of risky cash flows.

3.2.1 Misallocations and Aggregate Markups

This section illustrates how markup dispersion across industries creates misallocations (in the spirit of Lerner [34]). For ease of exposition, we introduce two statistics of the cross-sectional markup distributions for the macro-economy in each state s:

$$\bar{M}_s = G_{1-\theta}\left(M_s\right), \tag{13}$$

$$\eta_s = \left(\frac{G_{-\theta}\left(M_s\right)}{G_{1-\theta}\left(M_s\right)}\right)^{\theta} \leq 1. \tag{14}$$

where $G_p\left(M_s\right) = \left(\int \alpha_s(z) M_s\left(z\right)^p dz\right)^{\frac{1}{p}}$ refers to the p-th order cross-sectional power mean of $M_s\left(z\right)$.[12] These statistics capture distinct elements of the cross-sectional markup distribution, and are jointly sufficient in describing the aggregate economy. The variable $\bar{M}_s$ captures the notion of aggregate market power, i.e., an appropriate average markup across industries. The variable η_s captures the (inverse of) dispersion of markups across industries. By Jensen's inequality, η_s is bounded above by one (obtained when all in-

[12]Notice that by construction $\int_0^1 \alpha_s(z)dz = 1$, so we interpret α as a weighting measure where each industry obtains a weight according to its relative productivity.

dustries charge the same markup) and is decreasing in the dispersion of markups.[13] The variable η_s can be interpreted as a measure of allocative production efficiency.

Lemma 1. *Given the functions M_s, α_s and $\bar{A}_s$, aggregate consumption, C_s, real income y_s, in state s are given by:*

$$C_s = y_s = \bar{A}_s \eta_s. \tag{15}$$

The fraction of real *income that is derived from labor income is given by:*

$$\omega_s = \frac{1}{\eta_s \bar{M}_s}. \tag{16}$$

The outcome in state s is Pareto efficient if $M_s(z) \equiv k_s$ for all z, so that $\eta_s = 1$.

From equation 15, aggregate consumption only depends on the exogenous aggregate shock $\bar{A}_s$ and allocative efficiency η_s implied by the markup distribution. As long as markups do not vary across industries in each state (i.e., $M_s(z) \equiv k_s$ for all z and s), the allocation of labor to industries is efficient so that aggregate consumption, i.e., potential output, is given by the aggregate shock $\bar{A}_s$. In all such economies, *relative* goods prices match the perfectly competitive and hence efficient outcome. Allocative efficient economies can only differ in terms of the decomposition of income, i.e., the fraction of income derived from labor ω_s and from firm profits, which are redistributed to the representative agent. An important benchmark case is the monopolistic economy, in which $M_s(z) = \frac{\theta}{\theta-1}$ and $\omega = \frac{\theta-1}{\theta}$.

3.2.2 Valuation

A fundamental insight of the consumption based asset pricing literature is that the rate used to discount future cash flows should be intimately related to the state of the economy, and specifically to aggregate consumption. The general implication is that cash flows received in bad states of the world will be worth more than cash flows received in good states, and thereby discounted at a lower rate. The discount factor is thus stochastic; it depends on the realization of future consumption.[14]

We assume that there is a complete market of Arrow-Debreu securities in zero net supply, in addition to the stocks of the firms. The time t value of a stochastic cash flow

[13] This follows from the fact that $G_p(\tilde{x}) > G_q(\tilde{x})$ for any non-degenerate random variable $\tilde{x}$ as long as $p > q$.

[14] For a more extensive discussion, see, e.g., Duffie [18], Campbell [10], and references therein.

received at $t+1$, $Q_{s_{t+1}}$, is then $\mathbb{E}[SDF_{t+1} \times Q_{s_{t+1}}]$, i.e., the value is the expectation of the future cash flows discounted with the stochastic discount factor, the SDF (also called the pricing kernel). With our utility specification, $SDF_{t+1} = \delta \left(\frac{C_{s_{t+1}}}{C_{s_t}}\right)^{-\gamma}$. Given our decomposition of aggregate consumption into a productivity and a misallocation component (15), it follows that the SDF can be written as

$$SDF_{t+1} = \delta \left(\frac{C_{s_{t+1}}}{C_{s_t}}\right)^{-\gamma} = \delta \left(\frac{\bar{A}_{s_{t+1}}}{\bar{A}_{s_t}}\right)^{-\gamma} \left(\frac{\eta_{s_{t+1}}}{\eta_{s_t}}\right)^{-\gamma}. \tag{17}$$

Since equilibrium profits of a firm at time t depend only on the state, s, the information about the firm's future profits can be summarized in an S-vector, π, where π_s is the profit in state s. We also define the S-vector V, where V_s represents the current value of the firm if the current state is s. This value is the discounted value of a perpetuity of stochastic cash flows beginning in the next period.

Because of the Markovian structure of the state space (1), we have $P(s_{t+k} = j | s_t = i) = [\Phi^k]_{i,j}$, $k \geq 0$. The time-0 value of an Arrow-Debreu security that pays one Dollar at time t in state j, given that $s_0 = i$, is therefore $AD^t_{ij} = \delta^t \frac{C_j^{-\gamma}}{C_i^{-\gamma}} [\Phi^t]_{ij}$. We define the diagonal matrix Λ_m with its sth diagonal element made up by the marginal utility in state s, $[\Lambda_m]_{ss} = m_s = C_s^{-\gamma}$, and we can then write the value as $AD^t_{ij} = \delta^t [\Lambda_m^{-1} \Phi^t \Lambda_m]_{ij}$.

Using the Arrow-Debreu security prices, period-by-period and state-by-state, we obtain:

Lemma 2. *The state-contingent valuation V of a stochastic profit stream π is given by:*

$$V = \left[\Lambda_m^{-1}(I - \delta\Phi)^{-1}\Lambda_m - I\right] \pi. \tag{18}$$

This pricing formula differs from a risk neutral economy, in which there would be no marginal utility terms Λ_m (or, equivalently, it would be the case that $\Lambda_m = I$). The term Λ_m summarizes how valuations—and thereby the decisions of firms—are affected by risk aversion (through γ), aggregate productivity shocks (through $\bar{A}$), and misallocation (through η).

3.3 Industry equilibrium

Understanding strategic price setting behavior in one industry z is the first step towards endogenizing the entire markups function M. We therefore characterize, as a function of

industry and aggregate characteristics, when firms in a specific industry behave competitively, when monopolistic markups can be sustained, and when the outcome is neither of these extremes. Since each industry is small compared with the aggregate economy, firms in industry z take the dynamics of all other industries as exogenously given, i.e., they take M as exogenously given for all $z' \neq z$. In particular, the $S \times 2$ matrix consisting of the vectors C and $\bar{M}$ are *jointly sufficient* in describing the economic environment for one particular industry.

It is helpful to write real firm profits in sector z as a function of the choice variable $M_s(z)$ and the exogenous variables $C, \bar{M}$ and $\alpha(z)$. The expression follows directly from Lemma 1:

$$\pi_s(z) = \alpha_s(z) C_s \bar{M}_s^{\theta-1} \frac{M_s(z) - 1}{M_s(z)^{\theta}}. \tag{19}$$

While C_s and $\bar{M}_s$ are macro variables and hence affect all industries in a systematic fashion, the industry-specific productivity shock $\alpha_s(z)$ affects by definition only industry z. Note that industry z profits depend positively on the aggregate market power $\bar{M}_s$ since goods are substitutable (with $\theta > 1$).

In each state, s, firms in an industry choose the vector of state contingent markups to maximize the value function, $V_s(z)$, given the value maximizing behavior in each of the other states of the world, $V_{-s}(z)$, and subject to incentive compatibility ($\frac{V_s + \pi_s}{N(z)} \geq \pi_s$),

$$V_s(z) = \arg\max_{M_s} V_s(z) \,|V_{-s}(z), \tag{20}$$

for all s. Here, M_s maps to V_s via (18, 19).

Within our model's setting, finding the solution to the optimization problem (20) is straightforward by exploiting the linearity of the objective function and the constraints in profits. Since profits are not only affected by the choice of markups, but also the exogenous variables α, C and $\bar{M}$, we normalize profits (19) by monopoly profits $\pi_s^m(z)$:

$$\pi_s^N(z) \equiv \frac{\pi_s(z)}{\pi_s^m(z)} = \frac{M_s(z) - 1}{M_s(z)^{\theta}} \frac{(M^m)^{\theta}}{M^m - 1}. \tag{21}$$

This normalization provides a *state-independent* bijection with $\pi_s^N \leftrightarrow M_s$, where $1 \leq M_s \leq \frac{\theta}{\theta-1}$ and $0 \leq \pi_s^N \leq 1$. We also define the corresponding inverse function μ:

$$M_s(z) \equiv \mu\left(\pi_s^N(z)\right). \tag{22}$$

To capture the joint effect of the variables α, C and $\bar{M}$, it is useful to define a summary statistic of the severity of state-wise incentive constraints, $IC_s(z)$:

$$IC_s(z) = \alpha_s(z) C_s^{1-\gamma} \bar{M}_s^{\theta-1}. \tag{23}$$

Intuitively, $IC_s(z)$ consists of the state component of the current-period industry profit, $\alpha_s(z) C_s \bar{M}_s^{\theta-1}$, *weighted* by marginal utility $C_s^{-\gamma}$. The importance of this marginal utility effect is stronger the higher the risk-aversion coefficient γ. We collect $IC_s(z)$ in a diagonal matrix Λ_{IC}, so that the elements satisfy $[\Lambda_{IC(z)}]_{ss} = IC_s(z)$.

Using the definition of $IC(z)$ and $\pi^N(z)$, the dynamic equilibrium can now be viewed as a simple linear programming problem in which firms choose normalized profits $\pi_s^N(z) \leq 1$ instead of M_s in (20):

Proposition 1. *Given C and $\bar{M}$, the industry equilibrium outcome is uniquely determined by the solution to the following linear program.*

$$\pi^N(z) = \arg\max_{\hat{\pi}^N} \ \mathbf{1}^T \hat{\pi}^N, \quad s.t., \tag{24}$$

$$\hat{\pi}^N \leq \mathbf{1}, \tag{25}$$

$$0 \leq \left[(I - \delta\Phi)^{-1} - N(z) I\right] \Lambda_{IC(z)} \hat{\pi}^N. \tag{26}$$

The corresponding equilibrium markups satisfy $M_s(z) = \mu\left(\pi_s^N(z)\right)$. Unless the incentive constraint (26) binds in state s, the monopolistic outcome obtains, $M_s = M^m$.

The specific form of the incentive constraint, $\frac{V_s + \pi_s}{N(z)} \geq \pi_s$ (and its matrix counterpart 26), implies economically that an increase in the markup in state s', relaxes the incentive problem in all other states $s \neq s'$ due to an increase in V_s.[15] Thus, the dynamic optimization (20) can be represented as a static, state independent, linear programming problem (see simple objective (24)). Inspection of the program reveals that the exogenous variables α_s, C_s and $\bar{M}_s$ *only* affect the incentive constraint via $IC_s(z)$, giving it a key role for the comparative statics analysis (see subsequent Proposition 3).[16]

Going forward, it will be important to understand when the incentive constraint binds, so equilibrium markups deviate from the monopoly markup in at least some state.

[15] Recall that Φ is irreducible, so state s' will be reached with positive probability, regardless of the initial state s.

[16] Note that the s-th element of the vector $\Lambda_{IC(z)} \hat{\pi}^N$ is simply given by: $\hat{\pi}_s^N IC_s(z)$.

From (26), monopoly markups, $M^m = \mu(1)$, are sustainable in *all states* if and only if

$$\left[(I - \delta\Phi)^{-1} - N(z)\, I\right] \Lambda_{IC(z)} \mathbf{1} \geq 0. \tag{27}$$

By rearranging (27) for $N(z)$, we obtain a closed form expression for the threshold number of firms an industry, $N^m(z)$, for which monopolistic markups are sustainable for all s

$$N^m(z) = \min_s \Lambda_{IC(z)}^{-1} (I - \delta\Phi)^{-1} \Lambda_{IC(z)} \mathbf{1} \tag{28}$$

Intuitively, while for a small number of firms $N(z) \leq N^m(z)$, the monopoly outcome is sustainable in all states, too many firms in one industry, $N(z) > N^c$, imply the competitive outcome in all states. Only in the intermediate region may markups vary across states. This intuition is formalized in the following Proposition.

Proposition 2. *Given aggregate consumption C and the average markup $\bar{M}$, the equilibrium outcome satisfies:*

Normalized Profits	*Markups*	
$\pi_s^N(z) = 1$	$M_s(z) = \frac{\theta}{\theta-1}$	*for* $N(z) \leq N^m(z)$,
$\pi_s^N(z) \in \left(\frac{\underline{IC}(z)}{IC_s(z)}, 1\right]$	$M_s(z) \in \left(\mu\left(\frac{\underline{IC}(z)}{IC_s(z)}\right), \frac{\theta}{\theta-1}\right]$	*for* $N(z) \in (N^m(z), N^c)$,
$\pi_s^N(z) = \frac{\underline{IC}(z)}{IC_s(z)}$	$M_s(z) = \mu\left(\frac{\underline{IC}(z)}{IC_s(z)}\right)$	*for* $N(z) = N^c$,
$\pi_s^N(z) = 0$	$M_s(z) = 1$	*for* $N(z) > N^c$.

where $\underline{IC}(z) = \min_s IC_s(z)$ *and* $N^c \stackrel{\text{def}}{=} \frac{1}{1-\delta}$.

Before highlighting the general implications of Proposition 2, it is useful to illustrate the different regions in a stylized example with $S = 3$ states: Assume that aggregate consumption across states is $C = (1, 1.25, 1.875)^T$ and that aggregate markups are competitive in all states, $\bar{M} = (1, 1, 1)^T$. The transition between states is i.i.d. with all states being equally likely. Preference parameters are given by $\delta = 0.9$, $\gamma = 2$, and $\theta = 3$. Consider now an industry that moves one-to-one with the aggregate, i.e., $\alpha(z) = (1, 1, 1)^T$.

Since this example only features variations of aggregate consumption, we obtain $IC_s(z) = C_s^{1-\gamma}$ implying that the incentive problem is most severe in state 1 as $IC_1 > IC_2 > IC_3$. We immediately obtain from Equation 28 that $N^m = 8$. Thus, monopoly markups of $M^m = \frac{\theta}{\theta-1} = \frac{3}{2}$ are sustainable in *all states* if the number of firms satisfies

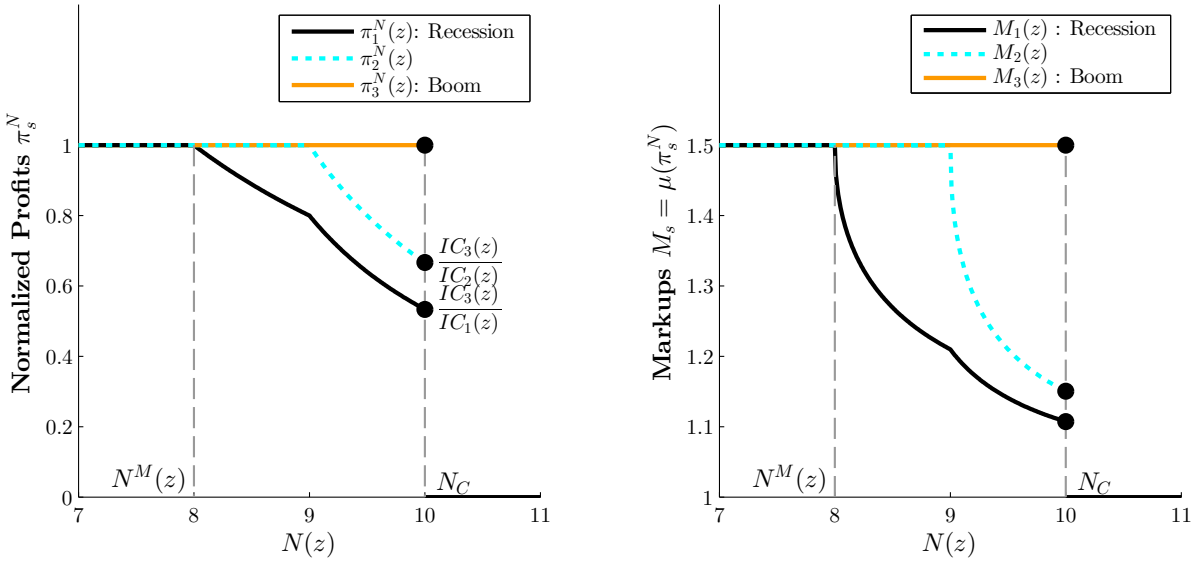

Figure 2. Left panel: This graph plots the state contingent normalized profits of one particular industry given aggregate consumption of $C = (1, 1.25, 1.875)^T$, aggregate markups of $\bar{M} = (1, 1, 1)^T$, and industry-specific shocks of $\alpha(z) = (1, 1, 1)^T$. We set $\gamma = 2$. If there are fewer than 8 firms in the industry, monopoly markups are sustainable in all states. Increasing the number of firms further causes the incentive constraint in state 1 to bind first, then in state 2 (at $N = 9$) and finally, at $N_C = 10$, all markups collapse discontinuously to the competitive outcome. **Right Panel:** The right panel plots the corresponding markups as a function of normalized profits.

$N \leq N_m = 8$. This can be directly inferred from Figure 2 which plots the optimal state-contingent normalized profits (left panel) and markups (right panel) as a function of the number of firms, confirming the four cases in Proposition 2. As soon as the number of players exceeds $N^m = 8$, the binding incentive constraint in state 1 pins down markups in state 1 while monopoly markups are initially still sustainable in states 2 and 3. When N exceeds 9 firms, monopoly markups can no longer be sustained in state 2 either. Interestingly, the binding incentive constraint in state 2 also has a (negative) feedback effect on the ability to collude in state 1 since the present value of *future* collusion profits is lowered in state 1, causing the kink in the state 1 markup function at $N(z) = 9$.[17] Finally, given $\delta = 0.9$, the threshold number of firms that induces the competitive outcome in all states is given by $N_c = 10$. This threshold $N^c = \frac{1}{1-\delta}$ only depends on the discount rate and is therefore independent of industry characteristics. The corresponding normalized profits are obtained in closed form: $\pi_s^N(z) = \frac{IC(z)}{IC_s(z)}$ so that $M_s(z) = \mu\left(\frac{IC(z)}{IC_s(z)}\right)$.[18]

We now return to our general analysis. From Proposition 2, one can immediately

[17] This graph implicitly treats N as a positive real number.

[18] It is easy to verify that for $\theta = 3$, we obtain that $\mu(x) = \frac{3}{\sqrt{x}} \cos\left(\frac{\arctan\sqrt{\frac{1-x}{x}} + \pi}{3}\right)$.

deduce that for all $N(z) \leq N^c$ there exists at least one state s in which the monopolistic outcome obtains, in particular the state(s) satisfying $IC_s(z) = \underline{IC}(z)$ (such as state 3 in the left panel of Figure 2). Intuitively, if there is no variation in incentive problems across states, i.e., $IC_s(z) = IC_{s'}(z)$ for all s, s', then $IC_s(z) = \underline{IC}(z)$ for all s and the monopoly outcome obtains in all states for $N(z) \leq N^m = N^c$ (and the competitive outcome obtains for $N(z) > N^c$). This insight leads to the following necessary conditions for markup variation across states:

Lemma 3. *Equilibrium markups may only vary across states if the following conditions are both satisfied:*
a) $IC_s(z) \neq IC_{s'}(z)$ *for some* s, s'*, and*
b) $N^m(z) < N(z) \leq N^c$.

If the intuitive conditions of a) time-varying incentive problems and b) intermediate competitiveness are satisfied, markup variation occurs on the equilibrium path. This motivates the following comparative statics analysis:

Proposition 3. *Equilibrium markups,* $M_s(z)$*, depend continuously on* C*,* $\bar{M}$*,* α *and* Φ.

1. *Equilibrium markups,* $M_s(z)$*, are decreasing in* $N(z)$ *for each* s.
2. *Equilibrium markups,* $M_s(z)$*, are decreasing in* $IC_{s'}(z)$ *for* $s = s'$ *and increasing in* $IC_{s'}(z)$ *for each* $s \neq s'$.

Continuity of markups is an important technical ingredient for the proof of Proposition 4. The intuitive, inverse relationship between markups and the number of firms N (Comparative static 1) can be immediately verified in the left panel of Figure 2. The comparative statics of $IC_s(z) = \alpha_s(z) \bar{M}_s^{\theta-1} C_s^{1-\gamma}$ represent a fundamental result of our analysis by relating the cyclicality of markups to α, $\bar{M}$ and C. An increase in $IC_s(z)$ will lower markups in that state (also compared to markups in other states $s' \neq s$). Since IC_s is increasing in α and $\bar{M}$, the comparative statics thus imply that markups are countercyclical with respect to the industry-specific component of profits α, and the average markup across all industries $\bar{M}$. Thus, markups exhibit strategic substitutability. Intuitively, when all industries charge on average a higher markup $\bar{M}_s$ in a given state, profits for a particular industry in that state will be higher since goods are substitutable (see 19). This increases the incentive to deviate, IC_s, and hence results in a lower equilibrium markup. Interestingly, the definition of $IC_s(z)$ implies that the dependency on aggregate consumption crucially depends on the risk aversion parameter γ, which we summarize in the following immediate corollary:

Corollary 1. *Aggregate consumption shocks and markup cycles:*

1. *If $\gamma < 1$, equilibrium markups, $M_s(z)$, are decreasing in $C_{s'}$ for $s = s'$ and increasing in $C_{s'}$ for each $s \neq s'$.*

2. *If $\gamma = 1$, equilibrium markups are independent of C_s for all s.*

3. *If $\gamma > 1$, equilibrium markups, $M_s(z)$, are increasing in $C_{s'}$ for $s = s'$ and decreasing in $C_{s'}$ for each $s \neq s'$.*

To understand why the threshold level for procyclicality is given by $\gamma = 1$, it is useful to separate out the forces of aggregate demand y_s and C_s in the definition of the summary statistic $IC_s(z)$, i.e., $IC_s(z) = \alpha_s(z) Y_s \bar{M}_s^{\theta-1} C_s^{-\gamma}$. Higher aggregate demand (c.p.) increases the temptation to deviate while lower marginal utility (higher C) reduces the incentive to deviate.[19] Since aggregate demand and consumption coincide in our framework, i.e., $Y_s = C_s$, the two forces exactly offset each other for $\gamma = 1$.[20] In the left panel of our example above, we set $\gamma = 2$ leading to procyclical markup variation when $N(z) \in (8, 10]$, i.e., $M_1(z) \leq M_2(z) \leq M_3(z)$. In Figure 3 we show the effect of varying γ (fixing N at N^c) on the cyclicality of markups.

Why does higher risk aversion, or equivalently lower $EIS = \frac{1}{\gamma}$, make it more attractive to deviate in bad times despite smaller profits? In bad times, the marginal value of consumption is higher causing today's valuations of future profits to be lower (see discussion in Section 3.2.2). Loosely speaking, when $\gamma > 1$ ($EIS < 1$) value-maximizing firms are (sufficiently) more desperate for an additional dollar in recessions. The marginal utility channel thus overturns the result of Rotemberg and Saloner [43].[21] Since misallocations through markup dispersion across industries feed back into the industry problem only via $C_s = A_s \eta_s$, the importance of this feedback effect relates to γ as well. For logarithmic utility ($\gamma = 1$), misallocations are thus irrelevant for the industry outcome.

[19] We thank an anonymous referee for suggesting this intuitive decomposition.

[20] While the exact threshold value for procyclicality of $\gamma = 1$ is a result of the equivalence of y_s and C_s, the general impact of discount rates qualitatively extend to setups when consumption and aggregate output are not identical, but positively correlated (such as in an economy with investment). In fact, when C is a linear function in y our results would apply one-to-one.

[21] Of course, if one considers specific aggregate shocks that purely affect y but do not (immediately) affect C, such as government expenditures in Rotemberg and Woodford [44], then markups are still countercyclical with respect to these shocks. However, in general, aggregate shocks both affect aggregate demand and consumption. Therefore, the cyclicality of markups relates to γ, see Corollary 1.

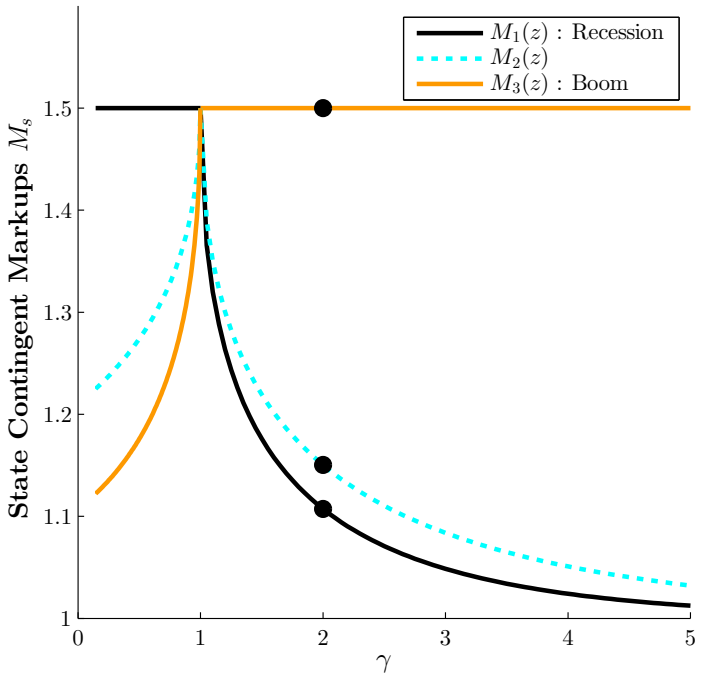

Figure 3. For the case $N(z) = N^c$, we plot the effect of the risk aversion parameter γ on the cyclicality of markups. Consumption is increasing in the state, s, $C = (1, 1.25, 1.875)^T$, and all other parameters are also as in Figure 2. Markups are countercyclical for $\gamma < 1$ (higher in states with lower consumption) and procyclical for $\gamma > 1$ (higher in states with higher consumption). The benchmark case of $\gamma = 2$ is highlighted with circles.

3.3.1 Endogenous entry

Before characterizing the general equilibrium implications, we briefly discuss the robustness of our results to endogenous entry. While our model technically does not allow for entry, let us now assume that an industry entrant faces a one-time entry cost of $\phi > 0$ to be able to enter the industry at $t+1$. Clearly, the decision of the entrant in an industry with currently $N^0(z)$ players depends on the assumed continuation equilibrium of that industry upon entry of an additional firm.

It is natural to select the equilibrium outcome described in Proposition 2, using $N(z) = N^0(z)+1$ as the post-entry equilibrium. Then, a firm will have a strict incentive to enter in state s if $V_s^{N^0(z)+1} > (N^0(z)+1)\phi$ where $V_s^{N^0(z)+1}$ is the present value of industry profits with $N^0(z)+1$ firms. Of course, a similar argument applies with $N^0(z)+1$ as the starting number of firms. Since the maximum number of firms in an industry is bounded above by N^c (due to positive entry cost), the decision of a potential

entrant in an industry with $N^0(z) = N^c - 1$ plays a special role by backward induction[22]: As long as $V_s^{N^c}(z) > N^c\phi$ for at least one state s, an industry will be populated by N^c firms in the long run. Thus, for sufficiently low entry cost the tightly characterized special case of $N(z) = N^c$ represents an economically meaningful outcome in an economy with endogenous entry.

Interestingly, in our model equilibrium profits and markups do not approach zero as the entry cost becomes arbitrarily small. This is because profits/markups are discontinuous at N^c (see Figure 2). We will revisit the important special case of $N = N^c$ in general equilibrium to obtain additional insights.

4 General Equilibrium

4.1 Existence and Uniqueness Conditions

We show the existence of general equilibrium in which firms in each industry choose optimal markups given the (optimal) markups chosen by firms in all other industries. Recall that the economy's environment is characterized by the tuple $\mathcal{E}$, i.e., by the real variables $\alpha : S \times [0,1] \to \mathbb{R}_+$, $N : [0,1] \to \mathbb{N}$, $g \geq 0$, $\bar{A} \in \mathbb{R}^S_{++}$, the irreducible aperiodic stochastic matrix, $\Phi \in \mathbb{R}^{S\times S}_{++}$, and the preference parameters, γ, θ, and δ. We note that a given equilibrium is completely characterized by the markup function, $M : S \times [0,1] \to \left[1, \frac{\theta}{\theta-1}\right]$, together with $\mathcal{E}$, since all other real and financial variables can be calculated from M and (13-19). This motivates the following

Definition 1. *General Equilibrium in economy $\mathcal{E}$ is given by a markup function $M : S \times [0,1] \to \left[1, \frac{\theta}{\theta-1}\right]$ for which,*

1. *$\bar{M}$ and C are defined by Equations 13 and 15,*
2. *For all z, $M(z)$ is the solution to the maximization problem given by Equations 24-26, with $M_s(z) = \mu\left(\pi_s^N(z)\right)$.*

We note that the existence and uniqueness of the second part of the definition is guaranteed by Proposition 1, industry by industry, i.e., given $\bar{M}$ and C there is a unique

[22]This follows from the fact that industry profits are decreasing in number of firms, which is a direct consequence of Proposition 3.

optimal markup function. It is a priori unclear, however, whether there exists a general equilibrium, i.e., whether both parts can be solved simultaneously. In other words, both the mappings, $M \mapsto (\bar{M}, C)$ (part 1) and $(\bar{M}, C) \mapsto M'$ (part 2) are well defined, but it is unclear whether M can be chosen such that the second step maps to the same markup function that was used in the first step, i.e., such that $M' = M$.

It turns out that we are able to prove the existence of equilibrium under very general conditions. Specifically, we assume that the functions N and α are Lebesgue measurable functions, and impose the following technical condition:

Condition 1. *For all s, for almost all z, $c_0 \leq \alpha_s(z) \leq c_1$ for constants, $0 < c_0 \leq c_1 < \infty$.*

We now have the following general result:

Proposition 4. *General equilibrium exists in any economy that satisfies Condition 1.*

Thus, only the technical conditions of integrability and boundedness of productivity functions across industries are needed to ensure the existence of equilibrium. The generality of this existence result is a priori quite surprising. In static general equilibrium models with imperfect competition, additional conditions in the form of quasi-concavity of firms' profit functions, and uniqueness of market clearing price functions given a productive allocation, are typically needed to show the existence of general equilibrium (see Gabszewicz and Vial [23]; Marschak and Selten [36]; and Benassy [5]). These conditions are indeed satisfied in our model, as seen in Section 2.1. Instead, the major challenge is the dynamic setting, where the move from a static to a dynamic Bertrand game between firms drastically enlarges the strategy space. Since all firms are intertwined through the effects their actions have on the pricing kernel, showing the existence under general conditions seems out of reach. Previous literature (e.g., Rotemberg and Woodford [44]; Gali [24]; and Schmitt-Grohe [46]) has avoided the issue by assuming complete symmetry, in which case the state space collapses. Of course, the focus on symmetric economies also restricts the type of effects that may arise, e.g., in terms of efficiency losses.

The reason why existence is still provable in our setting is the special structure of the model. The key property is that the game played between firms is simple enough that we can completely characterize their behavior under general parameter values and show that this behavior has some needed properties. Specifically, the structure of firms' constrained optimization problems in equations 24 - 26 allows us to show uniqueness and uniform continuity of industry outcomes with respect to all parameters. This follows from two

properties of the optimization problem. First, the objective function is linear. Second, the IC constraints have a specific form such that (i) for any number of firms less than the competitive threshold, $N < N^c$, the domain of optimization is uniformly bounded, closed, convex with nonempty interior, (ii) for industries with $N = N^c$ the domain is a closed bounded line, and (iii) for industries with $N > N^c$ the domain contains a single point, the origin. These properties imply well behaved (unique and uniformly continuous) outcomes industry-by-industry, which in turn implies that the mapping $M \mapsto (\bar{M}, C) \mapsto M'$ is continuous (in the function space L^1).

Technically, the proof of Proposition 4 depends on Schauder's fixed point theorem.[23] Specifically, it is shown in the proof of Proposition 4 that the space of markup functions is compact and convex, which, via Schauder's theorem, then guarantees the existence of a fixed point, i.e., an equilibrium. Details are given in the proof.

We note that Proposition 4 makes no claim as to equilibrium uniqueness. Uniqueness of equilibria can, however, be proved for the important benchmark case of homogeneous industries.

Proposition 5. *If industries in the economy $\mathcal{E}$ are homogeneous, i.e., if $a_s(z) \equiv 1$, for all z and s, and $N(z) \equiv N$ for all z, then the equilibrium is unique.*

Thus, if $N(z) \equiv N$ and each industry moves one-to-one with the aggregate shock $\bar{A}$, the industry outcome must not only be identical across industries, $M_s(z) = \bar{M}_s$, but $\bar{M}_s$ is also unique. Note that uniqueness and Pareto optimality of aggregate consumption, $C_s = \bar{A}_s$, follow directly from the lack of markup dispersion across industries (see Proposition 1).

Using the special case of homogeneous industries also allows us to cleanly illustrate that our result concerning pro- versus countercyclicality of markups is independent of misallocation and survives in general equilibrium. To make this result particular transparent, let us again consider the special case of $N(z) = N^c$. In general equilibrium, the previously exogenous average markup across industries, $\bar{M}_s$, is now endogenous. We obtain a simple, closed-form expression for the resulting general equilibrium markups:

Lemma 4. *If industries are homogeneous and $N(z) = N^c$, then markups in each industry are given by:*

$$M_s(z) = \bar{M}_s = \frac{\theta}{\theta - \frac{\bar{A}_s^{\gamma-1}}{\max_j\left(\bar{A}_j^{\gamma-1}\right)}} \tag{29}$$

[23] We use this theorem because we have a continuum of industries.

If $\gamma > 1$, markups are procyclical. If $\gamma < 1$, markups are countercyclical.

While the remaining results of the paper will be concerned about misallocation arising from differential behavior of industries, the detailed analysis of homogeneous industries proved useful by highlighting that the cyclicality of markups is unrelated to heterogeneity. We now turn to the question of how misallocation dynamics can arise endogenously when we depart from the homogeneity assumption.

4.2 Endogenous Misallocation Dynamics

What drives misallocation and misallocation dynamics? We know from the previous section that the realistic feature of industry heterogeneity must play an important role. While our rich framework allows us to introduce heterogeneity in terms of shock exposures $\alpha(z)$ and the number of firms $N(z)$ across a continuum of industries for an arbitrary number of states, we want to present simple, stylized examples to highlight the economic intuition. It is important to emphasize that the chosen parametrizations should therefore not be interpreted as real world calibrations of our framework. In the first example, presented in Section 4.2.1, we show how industry-specific shocks can be transmitted to the aggregate economy. Subsequently, Section 4.2.2 shows that small technological shocks may be amplified through feedback effects from the strategic behavior of other industries. Indeed, these feedback effects are sufficiently strong to generate multiplicity of equilibria.

4.2.1 Transmission of industry-specific shocks

We first consider an example without aggregate shocks that departs from homogeneity in the simplest possible way. There are two different types of industries, $j \in \{1, 2\}$ such that all industries $z \in I_j$ share the same industry-specific shocks α (see Table 1). Half of the industries are of type 1 and half of the industries are of type 2. Low entry cost in all sectors of the economy ensure that $N(z) = N^c$ for all z. Thus, the only source of heterogeneity results from industry-specific shocks: By construction, industry-specific shocks "average out" *across* industries, state by state. However, we note that the optimal choice of markups is influenced by differential incentives to deviate *across* states, industry by industry. As a result, intertemporal incentive constraints do not "average out" and misallocation may arise. Since $N(z) = N^c$, Proposition 2 implies that the equilibrium outcome for industry I_j given the macro states is $\pi_s^N(I_j) = \frac{IC(I_j)}{\alpha_s(I_j)\eta_s^{1-\gamma}\bar{M}_s^{\theta-1}}$ using $\bar{A}_s = 1$.

Type, j	I_j	$\alpha_1(I_j)$	$\alpha_2(I_j)$
1	$z \in [0, 0.5)$	1	$\frac{1}{2}$
2	$z \in [0.5, 1]$	1	$\frac{3}{2}$
$\bar{A}$		$\bar{A}_1 = 1$	$\bar{A}_2 = 1$

Table 1. Economy with two industries and two states.

Consider now the benchmark case of log utility ($\gamma = 1$) so that the feedback channel via misallocation η_s is shut down (see Corollary 1). Firms in industry 1 have the highest incentive to deviate in state 1 since the industry-specific shock in state 1, $\alpha_1(I_1) = 1$, is twice as high as compared to state 2, $\alpha_2(I_1) = \frac{1}{2}$. Therefore, $M_1(I_1) < M_2(I_1) = M^m$. By the same rationale, industry 2 features low markups in state 2, $M_2(I_2) < M^m$, and monopolistic markups in state 1. The resulting markup dispersion across industries in state 1 and state 2 implies the first immediate result: Industry-specific shocks alone can lead to misallocation.

Using $\theta = 3$ (see e.g., Fernandez-Villaverde et al. [21]), we obtain the following, *unique* equilibrium outcome, which is independent of the transition matrix ϕ (as $N = N^c$):

Outcomes	$s = 1$	$s = 2$
$M(I_1)$	1.09	1.5
$M(I_2)$	1.5	1.16
$C = \eta$	**0.966**	**0.986**

(30)

In this example, strategic industry behavior does not only cause inefficiencies via the channel of misallocation, but misallocation is also time varying, i.e., $\eta_1 < \eta_2 < 1$. An economy without aggregate shocks, $\bar{A}_1 = \bar{A}_2 = 1$, now features endogenous volatility with a 2% difference in aggregate consumption.

What causes higher dispersion of markups in state 1 than in state 2? To get the intuition behind this asymmetry, observe that the structure of the industry-specific shocks implies that industry 1 faces a higher (a 2 : 1) incentive to deviate in state 1, whereas industry 2 faces a lower (a 3 : 2) incentive to deviate in state 2. In equilibrium, this asymmetry is reflected in lower markups in industry 1 in the state where the deviation temptations are largest; creating higher dispersion between the low markup in industry 1 and the monopoly markups of industry 2 in state 1.[24] Thus, the state with higher disper-

[24]A second channel is given by the α weights in calculating dispersion. While both industries are equally weighted in state 1, industry 2 has an effective weight of $\frac{3}{4}$ in state 2, mechanically creating less

sion of industry-specific shocks, state 2, actually features lower dispersion of equilibrium markups. The example shows that misallocation results from a subtle mechanism, namely how the industries' ability to "collude across states" varies across industries.

By setting $\gamma = 1$, the example so far deliberately shuts down the feedback effect of misallocation into the industry optimization problem. We now consider the feedback effects of misallocation as we increase γ above 1. In asset pricing, values of γ between 1 and 10 are considered reasonable (see Mehra and Prescott [37]). By Corollary 1, higher risk aversion will facilitate collusion in the state with high aggregate consumption, i.e., state 2, and increase the incentive to deviate in state 1. This in turn creates downward pressure on markups of industry 1 in state 1 and will allow industry 2 to sustain markups above 1.16 in state 2. This adjustment of markups aggravates the differences in aggregate consumption by causing higher dispersion in state 1 and higher efficiency in state 2, leading to further feedback. The higher γ, the stronger the feedback effects in general equilibrium. When $\gamma = 10$, consumption values in state 1 and state 2 of the unique equilibrium are given by 0.963 and 0.99, thus raising the consumption difference across states by 36% relative to the benchmark case of logarithmic utility.

Finally, the just presented example allows us to highlight that the equilibrium outcome at the threshold level of $N^c = \frac{1}{1-\delta}$ is extremely sensitive to (unexpected) changes in the discount rate: An arbitrarily small decrease in the discount factor δ will take the economy from the unique collusive outcome to the unique efficient, competitive outcome with $C_1 = C_2 = 1$, regardless of γ.

4.2.2 Shock Amplification

The previous example revealed how purely industry-specific shocks are transmitted to the aggregate economy, leading to sizeable aggregate fluctuations. We now study an example in which small aggregate shocks are amplified through the feedback effect via the stochastic discount factor. We will choose the parametrization in such a way that these feedback effects are so strong that multiple equilibria arise.

In particular, consider the economy described in Table 2, with three distinct types of industries, I_1, I_2 and I_3, and $S = 2$ states. Thus, there is one very small industry (I_1), one large industry (I_2), and one medium-sized industry (I_3). The first two industries have many firms, $N = 19$, but they will still not be perfectly competitive, since $N^c = \frac{1}{1-\delta} = 20$.

dispersion.

Type, j	I_j	N	A_1	A_2	α_1	α_2
1	$z \in [0, 0.02)$	19	0.25	1	0.8728	1
2	$z \in [0.02, 0.81)$	19	1	1	1.0026	1
3	$z \in [0.81, 1]$	1	1	1	1.0026	1
$\bar{A}$					$\bar{A}_1 = 0.974$	$\bar{A}_2 = 1$

$$\Phi = \begin{bmatrix} 0.7 & 0.3 \\ 0.3 & 0.7 \end{bmatrix}$$

$$\gamma = 6, \qquad \theta = 1.1, \qquad \delta = 0.95.$$

Table 2. Economy with three industries and two states.

The third industry is monopolistic so that it will charge the markup $\frac{\theta}{\theta-1}$ regardless of the behavior in the first two industries. Columns 4 and 5 in Table 2 describe the absolute productivity shocks, A, in the two states. We see that only the very small first industry experiences any variation in productivity across the two states. The aggregate variation in productivity will therefore be small. In columns 6 and 7, we show the decomposition of the absolute productivity shocks into industry-specific and aggregate components, α and $\bar{A}$ (see equations 11 and 12).[25] The effect on aggregate productivity of the first industry's shock is about 2.5%, since aggregate productivity is 0.974 in the low-productivity state and 1 in the high-productivity state. This would also be the aggregate consumption in the two states in an efficient equilibrium.

Before analyzing the equilibrium in this economy, it is instructive as a reference case to study the economy which is identical to that in Table 2, except for that $A_1 = 1$ in industry 1. Hence, this is an economy with no productivity shocks, neither industry-specific nor aggregate, and it follows that $\bar{A}_1 = \bar{A}_2 = 1$ and $\alpha_s(z) \equiv 1$ in this reference economy. One easily verifies that the monopolistic outcome, in which markups $M \equiv \frac{\theta}{\theta-1} = 11$ are chosen by all firms in all states, is feasible in this case (this also follows as a consequence from Lemma 3), leading to the efficient outcome where $C_1 = \bar{A}_1 = 1$, $C_2 = \bar{A}_2 = 1$.

However, the efficient outcome cannot be sustained as an equilibrium in an economy with small productivity shocks. Instead, the following markup choices constitute an

[25] Note that the shock to industry 1 also affects the relative productivity in industries 2 and 3, since α is normalized to sum to one across industries, state by state.

equilibrium outcome

Equilibrium 1

Markups	$s=1$	$s=2$
$M(I_1)$	1.493	11
$M(I_2)$	1.4	11
$M(I_3)$	11	11
C	**0.782**	**1**

(31)

Thus, the small aggregate productivity shock ($\approx 2.5\%$) leads to a significant decrease in equilibrium output ($\approx 22\%$) in state 1. The intuition for why amplification occurs in this example is exactly in line with our main theme in this paper, that technological shocks which are small in aggregate — in that they only affect a few industries — change the strategic behavior of firms in other industries through the effect they have on the pricing kernel.

This mechanism is explained in Figure 4, focusing on the behaviors of industries 1 and 2.[26] In the upper part of the figure, the reference economy with identical industries is shown, in which case monopolistic profits are feasible for both industries, i.e., normalized profits $\pi_s^N(I_j) = 1$ for both industries and states. In the lower part of the figure, the economy in Table 2 is shown. Line A shows the relevant IC constraint in state 1, given the pricing kernel in the monopolistic outcome. Monopolistic profits are indeed feasible in industry 1 (lower left figure), but infeasible in industry 2 (lower right figure). Thus, the lower productivity in industry 1, through its effect on the pricing kernel, affects the outcome in sector 2, which moves the IC constraint in state 1 to line B. This in turn changes the pricing kernel even further, making monopolistic profits in industry 1 infeasible and further changing the outcome in industry 2, moving to lines C in the two industries, and generating further feedback effects. The ultimate effect of this mechanism is that the equilibrium moves to line D in the two figures, substantially different from the monopolistic equilibrium in the reference economy.

We just highlighted the important role of feedback effects via the stochastic discount factor for equilibrium behavior of industries. Indeed, the feedback effects can be so strong

[26]Industry 3 is always monopolistic. The reason that it is still important for the example is that substantial efficiency losses only occur when there is high variability in markups across sectors. If industry 3 was not present then the economy would always be close to efficient, since markups would be the same for the vast majority of industries in each state — almost identical to the markups charged in industry 2. In contrast, when industry 3 is present and industry 2 charges low markups, efficiency will be low.

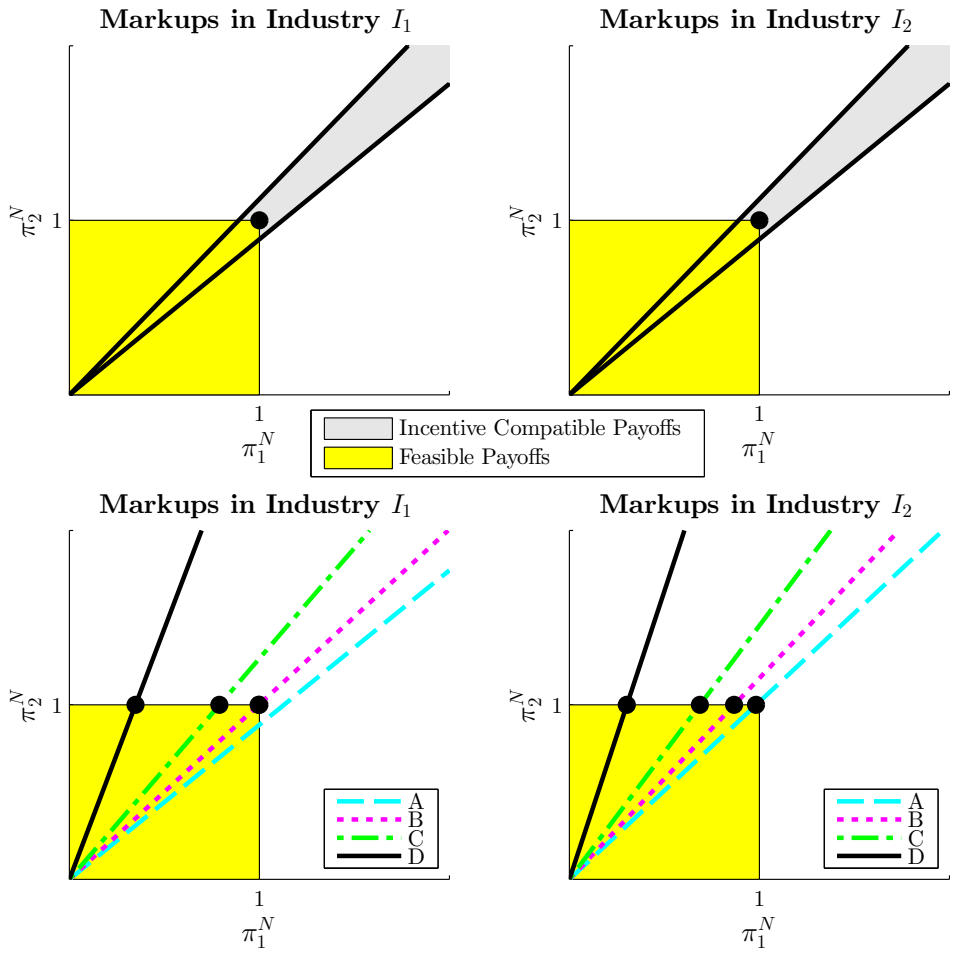

Figure 4. In each of the 4 panels, we plot incentive compatible and feasible normalized profits in both states of the world. Feasibility refers to the upper bound imposed by monopoly profits in each state, i.e., $\pi_s^N \leq 1$. Incentive compatibility in both states is governed by two lines. The upper line refers to the IC constraint in state 2. The lower one refers to the IC constraint in state 1. The upper 2 panels refer to the benchmark economy with identical industries. The outcome in industry 1 (2) is plotted on the left (right). In both industries and states monopolistic profits are sustainable. Below, we only plot the relevant IC constraint in state 1. Monopolistic profits violate IC constraint in state 1 for industry 2 (line A), in turn changing the IC constraints in state 1 for industry 1 (line B). The resulting equilibrium (line D) is substantially different.

that another equilibrium is consistent with the optimizing behavior of all industries. It can be verified that the heterogeneous economy $\mathcal{E}$ parameterized in Table 2 exhibits

(exactly) one more equilibrium supported by the following markups:

Equilibrium 2

Markups	$s=1$	$s=2$
$M(I_1)$	11	1.69
$M(I_2)$	11	1.904
$M(I_3)$	11	11
C	**0.974**	**0.841**

Again, aggregate fluctuations are endogenously determined. However, the second equilibrium is very different from the first one. First, although state 1 is the state that experiences the negative aggregate productivity shock, aggregate output is lower in state 2 due to the high dispersion of markups across industries (causing misallocation). Thus, there is a second way to ensure that firms do not deviate from equilibrium strategies, namely to decrease the attractiveness of state 2. A drop in consumption in state 2 via misallocation makes deviation more attractive in that state since $\gamma > 1$. This causes industries 1 and 2 to lower their markups in state 2 compared to the monopolistic industry 3, implying high markup dispersion which sustains the equilibrium outcome.

We note that multiplicity of equilibria is not a generic feature of our framework, but instead requires parametrizations that allow feedback effects of misallocation to be strong. The dual task of this stylized example to feature multiplicity and amplification, requires γ to be high (see Corollary 1) and θ to be low, making large variation between equilibrium markups across industries possible. However, we want to note that we can obtain significant amplification even in large-scale settings with standard parameters such as $\theta = 3$, (see e.g., Fernandez-Villaverde et al. [21]) and $\gamma = 3$.[27] Equilibrium multiplicity in models with markups, in the form of stationary sunspot equilibria, have also been generated in Gali [24] and Schmitt-Grohe [46]. The analysis in Gali [24] especially has similarities to ours in that he assumes linear production technologies and also covers the case with inelastic labor supply. However, his mechanism is different from ours. Since he focuses on the symmetric case with monopolistic competition, there is no role for heterogeneity in markups across firms, and the corresponding inefficiencies that such heterogeneity creates. Instead the multiplicity of equilibria arises because of self-fulfilling

[27] In a numerical exercise, we calculate the equilibrium outcome of 1,000 economies with 10,000 industries each. For each economy, industry-specific shocks and the number of firms were randomly generated. The assumed distributions of technology shocks implied a small aggregate productivity shock of $\Delta A = \left|\bar{A}_2 - \bar{A}_1\right| \approx 0.22\%$ across the 1000 simulations, resulting in a large shock to GDP/consumption, $\Delta C = |C_2 - C_1 \approx 0.82\%$. The GDP shocks are therefore on average about 3.8 times larger than the productivity shocks.

expectations about future growth rates. In our setup, industry homogeneity implies uniqueness as it prohibits efficiency losses due to cross-sectional variation of markups and hence shuts down the feedback channel through the pricing kernel.

5 Empirical Implications

Our theory has testable empirical implications for markup-cyclicality, for industries' joint effect on aggregate efficiency and economic activity, and for how strategic interaction at the industry level affects these aggregate variables in general equilibrium. While a rigorous empirical examination is beyond the scope of this theory paper, we summarize these implications, to provide a basis for future research.

First, our model has implications for the cyclicality of markups. To highlight the intuition for the sources of markup cyclicality and the sources thereof, it is useful to analyze the special case of an industry at the competitive threshold, i.e., $N(z) = N^c$, which should apply for any industry with small entry cost. Proposition 2 then implies the following structural expression for markup changes using $M_s(z) = \mu\left(\pi^N(z)\right)$.

$$\Delta \log \pi^N(z) = -\Delta \log \alpha(z) - (\theta - 1)\Delta \log \bar{M} + (\gamma - 1)\Delta \log(C) \tag{32}$$

where $\Delta \log x$ stands for $\log x_s - \log x_{s'}$. One may empirically estimate the relevance of the three factors presented in (32) via a standard time-series regression, industry by industry. First, markups should be countercyclical with respect to the industry-specific shock component α. To estimate the coefficient correctly, it would be important to use a raw measure of underlying demand/productivity shocks, i.e., a measure that is not contaminated by the endogenous markup choice (the left hand side measure). Second, markups are negatively related to average markups across industries since goods are substitutes. Third, the coefficient on the aggregate shock C depends on the risk aversion parameter γ. While our model features equivalence of Y and C, it may be empirically reasonable to both include Y, the aggregate demand channel, and C, the marginal utility channel, to account for the discrepancy between the two quantities in the data.

Second, in the aggregate, our model relates variation in economic activity to variations in allocative efficiency and technological shocks, i.e., $C_t = A_t \eta_t$. Since empirical studies are mostly concerned with growth, it is useful to express this identity as:

$$\Delta c = \Delta a + \Delta e, \tag{33}$$

where $c_t = \log(C_t)$, $a = \log(\bar{A})$, $e = \log(\eta)$, and Δ refers to first differences. From this expression, it is immediately clear that amplification of technological shocks, i.e., greater consumption volatility than suggested by technological condition ($\sigma_{\Delta c} > \sigma_{\Delta a}$), occurs if and only if

$$\rho_{\Delta a \Delta e} > -\frac{1}{2}\frac{\sigma_{\Delta e}}{\sigma_{\Delta a}}, \tag{34}$$

where $\rho_{\Delta a \Delta e}$ measures the coefficient of correlation between Δa and Δe. As a result, two factors can give rise to amplification: a high variation in efficiency relative to the variation in productivity ($\frac{\sigma_{\Delta e}}{\sigma_{\Delta a}}$) or a high positive correlation between efficiency and productivity ($\rho_{\Delta a \Delta e}$), i.e., countercyclical dispersion of markups. Both of these factors are quite intuitive. The relation allows one to estimate the importance of industry dynamics for aggregate fluctuations in the economy.

Finally, our model highlights the important role of industry characteristics capturing strategic interaction, such as Herfindahl indexes across industries and the dispersion of industry-specific shocks, for understanding the general equilibrium relationship between industry structure, markup variations, and aggregate fluctuations. In alternative theories of markup cyclicality, such as sticky-price models, such "strategic" variables would be irrelevant. We stress that the relationship is far more complex than simply one where less competition in an economy always leads to higher efficiency losses, as discussed in Section 3.2.1. To assess whether strategic interaction at the industry level represents a quantitatively important source of aggregate fluctuations, it would be interesting to estimate our model structurally.

6 Concluding Remarks

Our objective has been to understand the aggregate effects of strategic interaction between firms at the industry level. To achieve this, we develop a dynamic general equilibrium model featuring a continuum of different industries, each of which comprises a finite number of firms. The framework is tractable, and the strategic interaction between firms in each industry is straightforward to characterize. We establish the existence of general equilibrium and establish dynamic properties of the economy including equilibrium markups, firm profits and aggregate consumption.

The central premise of our model is that firms, maximizing shareholder value, are not always price takers but can be price setters. High prices in an industry can be sustained if firms value the future flow of profits over any immediate increases in market share

garnered by undercutting. Of course, the rate at which future profits are discounted depends both on the representative agent's preferences and on the behavior of the aggregate economy. Specifically, the misallocation of resources that arises from the equilibrium cross-sectional dispersion of markups affects aggregate consumption and therefore the representative agent's valuation of future profits. This feedback effect between industry equilibrium and the macro economy is the central intuition in our paper.

The strategic interaction yields various general equilibrium effects that can be interpreted in light of the macro economy. Even in an economy with no aggregate uncertainty, if the relative productivity of various industries changes, so does their ability to sustain collusive outcomes. These changes can affect both the level and the volatility of aggregate consumption. It is worthwhile to highlight how the interaction between industry heterogeneity and oligopolistic competition is key for our main general equilibrium effects: With fully flexible prices, dispersion of markups across industries can only arise if industries endogenously choose different markups. In an economy with *homogeneous* industries as in Rotemberg and Woodford [44], *oligopolistic* competition must lead to identical markups across industries, precluding real effects via misallocation. Under *monopolistic* competition it is irrelevant whether industries are heterogeneous, since all industries charge the same, monopoly markup. Thus, incorporating industry heterogeneity into a general equilibrium framework with oligopolistic competition generates a rich set of novel predictions.

An interesting implication of our analysis is that the social cost of collusion may be different from that calculated based on the forgone consumer surplus in any particular industry. Indeed, a standard partial equilibrium calculation, by definition, does not incorporate any social costs associated with resource misallocation, aggregate fluctuations and the ensuring general equilibrium change in valuations across industries. Operationally, it would be difficult to incorporate such costs, however it does suggest that in many cases, the costs of tacit collusion may be higher than usually calculated and a *macroprudential* view of anti-trust provisions is called for. Vigorous anti-trust enforcement in only a subset of industries may actually be welfare-decreasing.

A potentially fruitful extension of our model would be to consider asset pricing implications. The subgame perfect industry equilibria that we characterize naturally pin down the future value of each firm's cash flows. This of course, is the unlevered equity value of the firm. With an appropriate calibration, one could generate the relationship between returns, industry characteristics and the macro economy. We hope to explore these relationships in future research.

Acknowledgements

We thank the editors Alessandro Pavan and Xavier Vives as well as two anonymous referees for numerous insightful suggestions. In addition, the paper significantly benefited from helpful comments by Kyle Bagwell, Engelbert Dockner, Willie Fuchs, Nicolae Gârleanu, Zhiguo He, Anders Karlsson, Michael Katz, Debbie Lucas, Hanno Lustig, Miguel Palacios, Andrés Rodríguez-Clare, Julio Rotemberg, Stephanie Schmitt-Grohe, and Robert Staiger. Research assistance by Michael Weber is gratefully acknowledged. We also appreciated the feedback from seminar participants at Aarhus University, CAPR 2014, Goethe University, MIT, University of California, Berkeley, University of Illinois, Urbana Champaign, University of Southern California, the Mitsui Finance Conference, the 2012 Jackson Hole Finance Conference, and the 2013 meetings of the AEA. We thank the Clausen Center for International Business and Policy at U.C. Berkeley Haas School of Business for financial support.

A Data

To compute the time series of misallocations, we require a panel data set with markups for a large number of industries (ideally all) in an economy. The requirement of a large cross-section of industries makes it impossible to use state-of-the-art estimation techniques for markups that work well for one particular industry. Instead, we make use of the standard NBER manufacturing productivity database by Bartelsman and Gray containing information on 459 industries between 1959 and 2009. We exclude 8 discontinued industries leaving us with 451 industries.[28] We use (average) price cost margins (see Aghion et al. [3]) as a proxy for markups. Thus, $pcm_t(z)$, the estimate for industry z at time t is calculated as follows:

$$pcm_t(z) = \log(1 + PCM_t(z)) = \log\left(1 + \frac{\text{Value added}_t(z) \;-\; \text{Payroll}_t(z)}{\text{Value of Shipment}_t\;(z)}\right) \tag{A.1}$$

While this proxy is subject to shortcomings, such as not differentiating between marginal and average costs, it represents a reasonable proxy for a large scale study such as ours.[29]

B Proofs

Proof of Lemma 1

As explained in Section 3.1 we focus on time-invariant economies, so that all variables are solely expressed as state-dependent. Using the expression for prices, $p_s(z) = M_s(z)\frac{w_s}{A_s(z)}$ and the definitions of $\alpha_s(z)$, $\bar{A}_s$ and $\bar{M}_s$ (see equations 11, 12, and 13), we can solve for nominal prices and the nominal wage rate via normalizing the price index $P_s = \left(\int_0^1 p_s(z)^{1-\theta}dz\right)^{\frac{1}{1-\theta}}$ to one. Thus,

$$w_s = \frac{\bar{A}_s}{\bar{M}_s}, \tag{B.1}$$

$$p_s(z) = \frac{M_s(z)}{\bar{M}_s}\alpha_s(z)^{\frac{1}{1-\theta}}. \tag{B.2}$$

Finally, plugging the demand function of each sector, $c_s(z)$ (see equation 4) into the profit function of each sector $\pi_s(z)$ (see equation 6) yields an expression for y_s via the aggregate budget constraint (see equation 5)

$$y_s = \bar{A}_s\eta_s, \tag{B.3}$$

[28] Our results are virtually equivalent when we include those industries until their year of discontinuation.

[29] The proxy is consistent with our theory as the production function is constant returns to scale in labor (see De Loecker [15]).

where we have used the expression for nominal wages and prices (see equations B.1 and B.2) and the definition of η_s (see equation 14). Since the price index is normalized to one, $C_s = y_s$. The fraction of income derived by labor income, $\omega_s = \frac{w_s}{y_s}$, is readily obtained via equations B.1 and B.3. Real profits follow immediately from 4, 6, B.1, B.2, and B.3.

Proof of Lemma 2

By the definition of Arrow-Debreu prices, period-by-period and state by state, we obtain that:

$$\begin{aligned} V &= \sum_{t=1}^{\infty} \delta^t \Lambda_m^{-1} \Phi^t \Lambda_m \pi \\ &= \Lambda_m^{-1} \left(\sum_{t=1}^{\infty} \delta^t \Phi^t \right) \Lambda_m \pi \\ &= \Lambda_m^{-1} \left(\sum_{t=0}^{\infty} \delta^t \Phi^t - I \right) \Lambda_m \pi \\ &= \Lambda_m^{-1} ((I - \delta\Phi)^{-1} - I) \Lambda_m \pi \\ &= (\Lambda_m^{-1} (I - \delta\Phi)^{-1} \Lambda_m - I) \pi. \end{aligned}$$

The valuation operator $(\Lambda_m^{-1}(I - \delta\Phi)^{-1}\Lambda_m - I)$ has strictly positive elements. This implies represents the fact that higher profits in some state s strictly increases the present value of future profits, $V_{s'}$, in all states $s' = 1, \ldots, S$. Recall that Φ is irreducible, so each state will be reached with positive probability, regardless of the initial state.

Proof of Proposition 1

The Proposition is a special case of the following general lemma.

Lemma 5. *Consider a strictly positive vector $\pi^m \in \mathbb{R}^S_{++}$, a strictly positive matrix $\Theta \in \mathbb{R}^{S\times S}_{++}$, and a scalar $n \in \mathbb{R}_{++}$. Then there is a unique $\xi \in \mathbb{R}^S_+$ so that for all strictly positive $b \in \mathbb{R}^S_{++}$,*

$$\begin{aligned} &\xi = \arg\max_x \; b^T x, \; s.t., \\ &x \leq \pi^m, \\ &0 \leq (\Theta - nI)\, x. \end{aligned} \tag{B.4}$$

For each s, the solution has either the first or the second constraint binding, i.e., for each s, $\xi_s = \pi^m_s$ or $n\xi_s = \Theta\xi_s$.

Proof: Let $x < y$ denote that $x \leq y$ and $x \neq y$. Also, define $z = x \vee y \in \mathbb{R}^S$, where $z_s = \max(x_s, y_s)$ for all s. Clearly, $x \leq x \vee y$, where the inequality is strict if there is an s such that $y_s > x_s$. Finally, define the set $K = \{x : 0 \leq x,\; x \leq \pi^*, nx \leq \Theta x\}$. Note that K is compact.

Now, there is a unique maximal element of K, that is, there is a unique $\xi \in K$, such that for all $x \in K$ and $x \neq \xi$, $\xi > x$. This follows by contradiction, because assume that there are two distinct maximal elements, y and x, then clearly $z = x \vee y$ is strictly larger that both x and y. Now, it is straightforward to show that $z \in K$. The only condition that is not immediate is that $\Theta z \geq nz$. However, this follows from $\Theta(x \vee y) \geq \Theta x \vee \Theta y \geq nx \vee ny = n(x \vee y) = nz$.

Now, since b is strictly positive, it is clear that ξ is indeed the unique solution to the optimization problem regardless of b. That one of the constraint is binding for each s also follows directly, because assume to the contrary that neither constraint is binding in some state s. Then ξ_s can be increased without violating either constraint in state s and, moreover, the constraints in all the other states will actually be relaxed, so such an increase is feasible. Further, since $b_s > 0$, it will also increase the objective function, contradicting the assumption that ξ is optimal.

In particular, Lemma 5 can be applied to the industry optimization problem (see 20) by specifying $\Theta = (\Lambda_m^{-1}(I - \delta\Phi)^{-1}\Lambda_m - I)$, $V_s = \iota_s^T \Theta \pi$ (see Lemma 2), $b = \Theta^T \iota_s$ and $n = N - 1$. Then, the incentive constraint (8) can be written as $(\Theta - nI)\,\pi = V + \pi - N\pi \geq 0$.

Finally, we restate the program in Lemma 5 in terms of normalized profits. Normalized profits satisfy:

$$\pi_s^N(z) = \frac{\pi_s(z)}{\pi_s^m(z)} = \frac{\pi_s(z)}{\alpha_s(z)\, C_s \bar{M}_s^{\theta-1} \frac{M^m-1}{(M^m)^\theta}} = \frac{C_s^{-\gamma}}{IC_s(z)} \frac{\pi_s(z)}{\frac{M^m-1}{(M^m)^\theta}}$$

or in vector form:

$$\pi^N(z) = \frac{(M^m)^\theta}{M^m - 1} \Lambda_m \Lambda_{IC}^{-1} \pi(z) \tag{B.5}$$

Since b in the objective B.4 is just required to be strictly positive, we choose $b = \mathbf{1}$ for simplicity. By construction, normalized profits are bounded above by 1, i.e., feasibility implies:

$$\pi^N(z) \leq \mathbf{1}$$

This yields constraint 25. To obtain constraint 26, we need to rewrite the incentive constraint $V + \pi - N\pi \geq 0$. First note, that (B.5) implies:

$$\pi(z) = \frac{M^m - 1}{(M^m)^\theta} \Lambda_m^{-1} \Lambda_{IC} \pi^N(z)$$

Then:

$$\begin{aligned} V + \pi - N\pi &= (\Lambda_m^{-1}(I - \delta\Phi)^{-1}\Lambda_m)\pi - N\pi \\ &= (\Lambda_m^{-1}(I - \delta\Phi)^{-1}\Lambda_m - NI)\pi \\ &= \frac{M^m - 1}{(M^m)^\theta}((I - \delta\Phi)^{-1}\Lambda_m - NI)\Lambda_m^{-1}\Lambda_{IC}\pi^N(z) \\ &= \frac{M^m - 1}{(M^m)^\theta}\Lambda_m^{-1}((I - \delta\Phi)^{-1} - NI)\Lambda_{IC}\pi^N(z) \end{aligned}$$

Without loss of generality we can premultiply the incentive constraint $V + \pi - N\pi \geq 0$ with $\Lambda_m \frac{(M^m)^\theta}{M^m-1}$ so that we obtain the constraint 26.

Proof of Proposition 2

In this proof, the variable x is proportional to normalized profits π^N. Let $K^*(N) \stackrel{\text{def}}{=} \{x : 0 \leq x, N\Lambda_{IC}x \leq (I - \delta\Phi)^{-1}\Lambda_{IC}x\}$. Now, $N\Lambda_{IC}x \leq (I - \delta\Phi)^{-1}\Lambda_{IC}x$ is equivalent to $Ny \leq (I - \delta\Phi)^{-1}y$, where $y = \Lambda_{IC}x \in \mathbb{R}_+^S$. We first show that $K^*(n) = \{0\}$ when $N > \frac{1}{1-\delta}$, which immediately implies that the only solution to the optimization problem in Proposition 1 is indeed the competitive outcome. Define the matrix norm $\|A\| = \sup_{x \in \mathbb{R}^S \setminus \{0\}} \frac{\|Ax\|}{\|x\|}$, where the l^1 vector norm $\|y\| = \sum_s |y_s|$ is used. Since Φ is a stochastic matrix,

$\|\Phi^i\| = 1$ for all i and using standard norm inequalities it therefore follows immediately that

$$\|(I-\delta\Phi)^{-1}\| = \left\|\sum_0^\infty \delta^i\Phi^i\right\| \leq \sum_0^\infty \delta^i\|\Phi^i\| = \frac{1}{1-\delta},$$

and thus $\|(I-\delta\Phi)^{-1}y\| \leq \frac{1}{1-\delta}\|y\|$. Now, $Ny \leq (I-\delta\Phi)^{-1}y$ implies that $N\|y\| \leq \|(I-\delta\Phi)^{-1}y\|$, and therefore it must be the case that $N \leq \frac{1}{1-\delta}$, for the inequality to be satisfied for a non-zero y. Now, consider the case when $N = \frac{1}{1-\delta}$. Since $y = \mathbf{1}$ is an eigenvector to Φ with unit eigenvalue, it is also an eigenvector to $(I-\delta\Phi)^{-1}$ with corresponding eigenvector $\frac{1}{1-\delta}$, leading to $x = \Lambda_{IC}^{-1}\mathbf{1}$ or $x_s = IC_s^{-1}$. It is easy to show that this is the unique (up to multiplication) nonzero solution. Given the properties of Φ, the Perron-Frobenius theorem implies that this is indeed the *only* eigenvector with unit eigenvalue, and therefore also the only eigenvector to $(I-\delta\Phi)^{-1}$ with eigenvalue $\frac{1}{1-\delta}$. Now, take an arbitrary $y \in \mathbb{R}_+^S \backslash \{0\}$ as a candidate vector to satisfy the inequality, i.e., such that $z = (I-\delta\Phi)^{-1}y$ satisfies $z_i \geq Ny_i = \frac{1}{1-\delta}y_i$ for all i. Then, since $\|(I-\delta\Phi)^{-1}\| = \frac{1}{1-\delta}$, it follows that $\sum_i z_i \leq \frac{1}{1-\delta}\sum_i y_i$. The two inequalities can only be satisfied jointly if $z_i = \frac{1}{1-\delta}y_i$ for all i, and thus y is the already identified eigenvector. Thus, $K^*\left(\frac{1}{1-\delta}\right) = \{\Lambda_{IC}^{-1}\mathbf{1}\sigma, \sigma \geq 0\}$. Since $\pi^N \propto x$ with the additional constraint $\pi^N \leq \mathbf{1}$, the maximal σ that satisfies $\sigma IC_s^{-1} \leq 1$ for all s is given by $\min_s IC_s$. This leads to normalized profits of $\pi^N = \frac{\min_s IC_s}{IC_s}$.

Proof of Lemma 3

The second statement $N^m(z) < N(z) \leq N^c$ follows directly from the discussion in the text. Second, we prove that IC_s must vary across states for markup variation to occur. If $IC_s = k$ for some constant k, the diagonal matrix Λ_{IC} becomes $\Lambda_{IC} = kI$ so that we obtain for $N^m(z)$ (see 28):

$$N^m(z) = \min_s \Lambda_{IC}^{-1}(I-\delta\Phi)^{-1}\Lambda_{IC}\mathbf{1} = \min_s (I-\delta\Phi)^{-1}\mathbf{1} = \frac{1}{1-\delta} = N^c.$$

This is because the eigenvalue of $(I-\delta\Phi)^{-1}$ associated with the eigenvector of $\mathbf{1}$ is given by $\frac{1}{1-\delta}$ (see Proof of Proposition 2). So, $N^m = N^c$. Hence markups can never differ across states.

Proof of Proposition 3

Continuity follows from the fact that the objective function in Lemma 5 is a continuous function of all parameters and that (as long as N is strictly below N^c) the set K (see Proof of Lemma 5) is compact, and depends continuously on all parameters, in the sense that if K and K' are defined for two sets of parameter values, then $D(K,K')$ approaches zero when the parameter values that define K' approach those that define K. Here, $D(K,K') = \sup_{x\in K'} \inf_{y\in K} |x-y|$.

(1) follows from the definition of K in the proof of Lemma 5. It immediately follows that the set K is decreasing in N, which in turn immediately implies (1).

To prove the comparative statics in IC_s, i.e., claim (2), rewrite the incentive constraint in the program of Proposition 1, i.e., $\left[(I-\delta\Phi)^{-1} - N(z)I\right]\Lambda_{IC(z)}\hat{\pi}^N \geq 0$ as $\Omega\Lambda_{IC}\hat{\pi}^N$. Let $\Omega_{i,j}$ denote the i,j element of $\Omega = (I-\delta\Phi)^{-1} - N(z)I$. Note that only the diagonal elements $\Omega_{s,s}$ may be negative. The s element of the vector $\Omega\Lambda_{IC}\hat{\pi}^N$ is simply:

$$v_s = \Omega_{s,s}IC_s\pi_s^N + \sum_{j\neq s}\Omega_{j,s}IC_j\pi_j^N \geq 0 \tag{B.6}$$

We first note that if the incentive constraint binds in some state s, i.e., $v_s = 0$, then this implies that $\Omega_{s,s} < 0$ since $\Omega_{j,s} > 0$ for all $j \neq s$ (see Proof of Lemma 2). We now consider the comparative statics as we change the k-element of IC by Δ_k. We denote the outcome of the (new) optimization problem $v_s(\Delta_k), \pi^N(\Delta_k)$.

Case 1: Suppose first that the incentive constraint does not bind in state k when $\Delta_k = 0$, i.e., $\pi_k^N(0) = 1$. Moreover, let Δ_k be sufficiently small, such that the constraint in state k is still slack after the increase in Δ_k, i.e., $\pi^N(\Delta_k) = 1$.[30] Then, the incentive constraint (B.6) in all states $s \neq k$ is relaxed by $\Omega_{k,s}\Delta_k 1 = \Omega_{k,s}\Delta_k$ (recall $\Omega_{k,s} > 0$ and $\pi_k^N(0) = 1$). Therefore, for any state with a previously binding incentive constraint, i.e., $v_s(0) = 0$, there is now a *strict* increase in the markup, i.e., $\pi_s^N(\Delta_k) > \pi_s^N(0)$ whereas π_k^N remains (by construction) unaffected.

Case 2: Suppose now, that the incentive constraint binds in some state k, i.e., $v_k = 0$ and hence $\pi^N(0) < 1$. Since the incentive constraint binds in state k, this implies that $\Omega_{k,k} < 0$ (for otherwise (B.6) cannot bind). Rearranging (B.6) implies:

$$\pi_k^N IC_k = \frac{\sum_{j \neq s} \Omega_{j,k} IC_j \pi_j^N}{|\Omega_{k,k}|} \tag{B.7}$$

Note, with $\Delta_k = 0$ it is impossible to find any incentive compatible way to increase the product $\pi_k^N(0) IC_k(0)$. This follows by definition of $\pi_k^N(0)$ being the maximum (and IC_k being a constant). Now, if we increase IC_k by Δ_k, then it also must be impossible to increase $\pi_k^N(\Delta_k) IC_k(\Delta_k)$. Suppose it was possible to increase $\pi_k^N IC_k$, then it would also be possible to find a $\pi^N > \pi^N(0)$ in an incentive compatible way when $\Delta_k = 0$. Contradiction. Thus, at best $\pi_k^N IC_k$ stays constant. If $\pi_k^N IC_k$ is held constant, markups in all other states are unaffected, i.e., $\pi_j^N(\Delta_k) = \pi_j^N(0)$. This can be trivially achieved by setting $\pi_k^N(\Delta_k)$ to

$$\pi_k^N(\Delta_k) = \pi_k^N(0) \frac{IC_k(0)}{IC_k(\Delta_k)} \tag{B.8}$$

Thus markups in state k are strictly decreasing in IC_k if the incentive constraint bind in state k. (all other markups are unaffected). By combining cases 1 and 2, we get the comparative statics in IC_s and IC_s'.

Proof of Proposition 4

Before showing existence, we discuss some invariance results which will be helpful in the proof. We first note that the following result follows immediately from Proposition 2:

Lemma 6. *In any general equilibrium, any two industries with the same N and α have the same markups, M, and profits, π.*

Also, we observe that it is only the distributional properties of N and α that are important for the aggregate characteristics of an equilibrium. This should come as no surprise given that the aggregate variables important for industry equilibrium only depend on the distributions. To be specific, we define the (cumulative) distribution function $F : \mathbb{N} \times [c_0, c_1]^S \to [0,1]$, where $F(n, s_1, \ldots, s_S) = \lambda(\{(z : N(z) \leq n : \wedge : \alpha_1(z) \leq s_1 : \wedge \cdots : \wedge : \alpha_S(z) \leq s_S\})$, and λ denotes Lebesgue measure. Thus, $F(n, \alpha_1, \ldots, \alpha_S)$ denotes the fraction of industries with number of firms less than or equal to n, and productivities $\alpha_s(z) \leq \alpha_s$ for all s. We say that two economies, $\mathcal{E}_1$ and $\mathcal{E}_2$, are equivalent in distribution if they have the same distribution functions, and agree on the other parameters: g, $\bar{A}$, Φ, γ, θ and $\hat{\delta}$. Also, two outcomes—in two different economies—are said to be equivalent if any two industries, z and z' in the first and second economy, respectively, for which $N^1(z) = N^2(z')$ and $\alpha_s^1(z) = \alpha_s^2(z')$ for all s, have the same industry markups in each state of the world, $M_s^1(z) = M_s^2(z')$ for all s.

[30] Thus, if $\Theta_{kk} < 0$, we require that $\Delta_k \Theta_{kk} + \varepsilon \geq 0$.

We then have

Lemma 7. *Given two economies that are equivalent in distribution. Then for each equilibrium in one of the economies there is an equivalent equilibrium in the other.*

We now prove the proposition with a fixed point argument, and therefore define a fixed point relationship for the markup function, M, which ensures that it defines an equilibrium. We define $R \stackrel{\text{def}}{=} \bar{N} \times [c_0, c_1]^S$, where $\bar{N} = \{1, 2, .., \lfloor N_c \rfloor + 1\}$, with elements $x = (n, \alpha_1, \ldots, \alpha_S) \in R$. We will then work with functions $M^0 : R \to [0,1]^S$, and given such a function, the transformation to the standard markup function is given by $M_s(z) = M_s^0(\min(N(z), \lfloor N_c \rfloor + 1), \alpha_1(z), \ldots, \alpha_S(z))$. The reason why we work with the canonical domain, R, rather than $S \times [0,1]$, is that compactness properties needed for a fixed point argument are easier obtained in this domain. Given a function, $M^0 : R \to \left[1, \frac{\theta}{\theta-1}\right]^S$, we define

$$p_s^0 = G_{-\theta}(M_s) = \left(\int \alpha_s(z) M_s(z)^{-\theta} dz \right)^{\frac{1}{-\theta}} = \left(\int_{x \in R} x_{s+1} M^0(x)^{-\theta} dF(x) \right)^{\frac{1}{-\theta}}, \tag{B.9}$$

$$p_s^1 = G_{1-\theta}(M_s) = \left(\int \alpha_s(z) M_s(z)^{1-\theta} dz \right)^{\frac{1}{1-\theta}} = \left(\int_{x \in R} x_{s+1} M^0(x)^{1-\theta} dF(x) \right)^{\frac{1}{1-\theta}}. \tag{B.10}$$

It follows immediately that the mapping from M^0 to p_0 and p_1 is continuous (in L^1 topology) and since $\int \alpha(z) dz = 1$, that p_s^0 and p_s^1 lie in $[1, \theta/(\theta-1)]$. From (15), it follows that

$$C_s = \bar{A}_s \left(\frac{p_s^0}{p_s^1} \right)^{\theta}, \tag{B.11}$$

and from (19) for $M(z) = \frac{\theta}{\theta-1}$ that

$$\pi_s^m = \frac{1}{p_1^{1-\theta}} \frac{(\theta-1)^{\theta-1}}{\theta^\theta} \alpha_s C_s = \frac{1}{p_1^{1-\theta}} \frac{(\theta-1)^{\theta-1}}{\theta^\theta} x_{s+1} C_s. \tag{B.12}$$

Now, for each z, given $\pi^m \in \mathbb{R}_+^S$, the program in Proposition 1 provides a continuous mapping from π^m to

$$\pi_s \in \prod_1^S [0, \pi_s^m]. \tag{B.13}$$

We use (19) to define the operator $\mathcal{F}$, which operates on functions, and which is given by:

$$M_s^1(x) = (\mathcal{F}(M^0)(x))_s = 1 + \frac{p_1(s)^{1-\theta}}{C_s x_{s+1}} (M_s^0(x))^\theta \pi_s.$$

Since each operation in (B.9-B.13) is continuous, it follows that $\mathcal{F}$ is a continuous operator (in $L^1(\mathbb{R}^{1+S})$-norm). Further, it also follows that if $M_s^0(x) \in \left[1, \frac{\theta}{\theta-1}\right]$, then since $0 \leq \pi \leq \pi^m$, $1 \leq M_s^1(x) \leq 1 + \frac{(\theta-1)^{\theta-1}}{\theta^\theta} (M_s^0)^\theta \leq \frac{\theta}{\theta-1}$. Define, Z as the set of all functions, $M : R \to [1, \theta/(\theta-1)]^S$, such that M is nonincreasing in its first argument and nondecreasing in all other arguments. Then, from what we have just shown, together with Proposition 3, it follows that $\mathcal{F}$ is a continuous operator that maps Z into itself. We also have

Lemma 8. *Z is convex and compact.*

We prove that the set, W, of nondecreasing functions $f : [0,1] \to [0,1]$, is convex and compact. The generalization to functions with arbitrary rectangular domains and ranges, $f : \prod_1^N [a_i, b_i] \to \prod_1^M [c_i, d_i]$, is straightforward, as is the generalization to functions that are nonincreasing in some coordinates and nondecreasing on others (as is Z). Convexity is immediate. For compactness, we show that every sequence of functions $f^n \in W$, $n = 1, 2, \ldots$, has a subsequence that converges to an element in W. First, note that W is closed, since a converging (Cauchy) sequence of nondecreasing functions necessarily converges to a nondecreasing function. To show compactness, define the corresponding sequence of vectors $g^n \in [0,1]^{2^j}$, for some $j \geq 1$, by $g_k^n = f_n(2^{-j}k)$, $k = 0, 1, \ldots 2^j - 1$. Now, since $[0,1]^{2^j}$ is compact it follows that there is a subsequence of $\{f^n\}$, $\{f^{n_m}\}$ that converges at each point $2^{-j}k$, to some $g^* \in [0,1]^{2^j}$. Define the function $h^j : [0,1] \to [0,1]$ by $h^j(x) = g_k^*$, for $2^{-j}k \leq x < 2^{-j}(k+1)$, which is obviously also in W. Next, take the sequence $\{f^{n_m}\}$, and use the same argument to find a subsequence that converges in each point $2^{-(j+1)}k$, $k = 0, \ldots, 2^{j+1} - 1$, and the corresponding function $h^{j+1}(x)$. By repeating this step, we obtain a sequence of functions in W, $h^j, h^{j+1}, \ldots$, such that for $m > j$,

$$\int_0^1 |h^m(x) - h^j(x)|dx \leq \sum_k (g_{k+1}^j - g_k^j)2^{-j} \leq 2^{-j}.$$

Thus, $h^j, h^{j+1}, \ldots$ forms a Cauchy-sequence, which consequently converges to some function $h^* \in W$. Take a subsequence of the original sequence of functions, $\{f^{n_j}\}$, such that $\int |f^{n_j} - h^j|dx \leq 2^{-j}$. Then, for $m > j$, since

$$\begin{aligned}
\int_0^1 |f^{n_m}(x) - f^{n_j}(x)|dx &= \int_0^1 |f^{n_m}(x) + h^m(x) - h^m(x) + h^j(x) - h^j(x) - f^{n_j}(x)|dx \\
&\leq \int_0^1 |f^{n_m}(x) - h^m(x)|dx + \int_0^1 |f^{n_j}(x) - h^j(x)|dx \\
&+ \int_0^1 |h^m(x) - h^j(x)|dx \\
&\leq 3 \times 2^{-j},
\end{aligned}$$

$\{f^{n_j}\}$ is also a Cauchy sequence and converges to $h^* \in W$. Thus, W is compact and the lemma is proved. Given Lemma 8 and the continuity of $\mathcal{F}$, a direct application of Schauder's fixed point theorem implies that there is a $M^* \in Z$, such that $\mathcal{F}(M^*) = M^*$. Now, given such a M^*, and its associated π^m defined by (B.12), and given the functions, $N(z)$ and $\alpha_s(z)$, $0 \leq z \leq 1$, Lemma 5 can be used to construct $M_s(z)$. Since M and M^* have the same distributional properties, and C, p_0 and p_1, only depend on distributional properties, it immediately follows that M constitutes an equilibrium. We are done.

Proof of Proposition 5

First note that an equivalent formulation of Lemma 5 is the following: Define the sets $\Xi_s = \{x \in \mathbb{R}_+^S : x_s \leq \pi_s^m\}$, $Q_s = \{x \in \mathbb{R}_+^S : 0 \leq ((\Theta - nI)\, x)_s\}$, where $\Theta = (\Lambda_m^{-1}(I - \delta\Phi)^{-1}\Lambda_m - I)$ and $n = N - 1$, and $R = (\cap_{s=1}^S \Xi_s) \cap (\cap_{s=1}^S Q_s)$. Then there is a unique element, $r \in R$, such that for all s, $r_s = \max_{q \in R} q_s$. That is, there is a unique element that jointly maximizes all coordinates of elements in R. Moreover, for each s, such that $r_s < \pi_s^m$ it must be that $r_s = \frac{1}{n}(\Theta x)_s$.

For coordinates such that $r_s < \pi_s^m$, if any number of the π_s^m is replaced by $\hat{\pi}_s^m > \pi_s^m$, i.e., if Ξ_s is replaced by $\hat{\Xi}_s = \{x \in \mathbb{R}_+^S : x_s \leq \hat{\pi}_s^m\}$, where $\hat{\pi}_s^m \geq \pi_s^m$, and the equality is only allowed to be strict for coordinates where $r_s < \pi_s^m$, and $\hat{R}$ is defined as $R = (\cap_{i=1}^S \hat{\Xi}_s) \cap (\cap_{i=1}^S Q_s)$, then $\hat{R} = R$, and consequently, $\hat{r} = r$ where $\hat{r}$ is the unique maximal element in $\hat{R}$. To see this, assume that an element $v \in \hat{R}$ existed such that $v_s > \pi_s^m$ for at least one s. Then since $\hat{R}$ is convex there must also be an element, $w = \lambda r + (1 - \lambda)v \in \hat{R}$, with $w_s \leq \pi_s^m$, for all s and $w_s = \pi_s^m$ for one coordinate such that

$r_s < \pi_s^m$. But then $w \in R$, and it must then be that $r_s = \pi_s^m$, leading to a contradiction. Thus, no such element exists, so $\hat{R} = R$.

Now, from our discussion in Section 3.3, it follows that in an equilibrium in a homogeneous economy, all firms must charge the same markups in any state, $M_s(z) = \bar{M}_s$ for all z, and that any equilibrium must be efficient so that $C_s = \bar{A}_s = A_s(z)$ and $\alpha_s(z), = 1$ for all s for all z. What is not a priori clear is whether there may be multiple average markup vectors, $\bar{M}$, that constitute an equilibrium. We now show that this is not the case.

Given an equilibrium in a homogeneous economy, it follows from equation (19), and that $C_s = A_s$, that

$$\frac{1}{\bar{M}_s} = 1 - \frac{\pi_s}{A_s} = (1 - u_s). \tag{B.14}$$

Here $u_s = \frac{\pi_s}{A_s} \in \left[0, \frac{1}{\theta}\right]$ represents firm profits in state s as a fraction of total output.

It further follows from $\pi_s^m \equiv \frac{M^m - 1}{(M^m)^\theta} C_s \alpha_s(z) \bar{M}_s^{\theta - 1}$ that given such an average markup across industries, the monopolistic profits as a fraction of total output in one (zero-measure) industry, z, that deviates from the average markup function is $\hat{u}_s = \frac{\hat{\pi}_s^m}{A_s} = \frac{M^m - 1}{(M^m)^\theta} \bar{M}_s^{\theta - 1} = \frac{M^m - 1}{(M^m)^\theta} (1 - u_s)^{1-\theta}$. We note that $\hat{u}_s \geq u$ for all $u \in \left[0, \frac{1}{\theta}\right]$, and that the inequality is strict except for at $\hat{u} = u = \frac{1}{\theta}$.

Given the homogeneous behavior of all other industries, the firm optimization problem in (24-26) can be written

$$\hat{u} = \arg\max_{\hat{u}} \ \iota_j^T \Lambda_A^{-1} \Theta \Lambda_A \hat{u}, \quad \text{s.t.,} \tag{B.15}$$

$$\hat{u}_s \leq \frac{M^m - 1}{(M^m)^\theta} (1 - u_s)^{1-\theta}, \quad s = 1, \ldots, S, \tag{B.16}$$

$$0 \leq \left(\Lambda_A^{-1} \Theta \Lambda_A - (N-1) I\right) \hat{u}, \tag{B.17}$$

where $\Lambda_A = diag(\bar{A}_1, \ldots, \bar{A}_s)$. A necessary and sufficient condition for u to be an equilibrium is now that $\hat{u} = u$ in the above optimization problem.

Assume that we have found such a u (we know that there exists at least one such u from the existence theorem). If we can show that u is also the solution to the same program, but where (B.16) is replaced by $\hat{u}_s \leq \frac{1}{\theta}$ for all s, then we are done, since there is a unique solution for that optimization problem (as follows from an identical argument as the proof of Lemma 5).

An identical argument as in Lemma 5 implies that for each s, either (B.16) or (B.17) binds (or both). For any s such that (B.16) binds, it must further be that equilibrium markups in that state are monopolistic, i.e., $u = \frac{1}{\theta}$. Thus, relaxing the constraints for those s to $\hat{u}_s \leq \frac{1}{\theta}$ does not change the solution to the problem.

For any other s, where (B.16) does not bind and (B.17) binds, we note that since $u_s < \frac{1}{\theta}$, $u_s < \frac{M^m - 1}{(M^m)^\theta} (1 - u_s)^{1-\theta}$, $\hat{u}_s$ is strictly lower than its bound imposed by (B.16) for such s. However, from the argument at the beginning of this lemma, it follows that relaxing the constraint for these coordinates does not change the solution, so we can relax the constraints to $\hat{u}_s \leq \frac{1}{\theta}$ for such s too. Thus, u is also a solution to the relaxed problem, and is therefore unique. We are done.

C Long Term Growth

When $g > 0$, we can still solve for time-invariant equilibria through appropriate normalizations. That is, we focus on equilibria which—except for the constant growth rate g—are time invariant in that outcomes

are the same at t_1 and t_2 if the states are the same, i.e., if $s_{t_1} = s_{t_2}$. In such equilibria, outcomes *on the equilibrium path* can be written as:

$$C(t) = (1+g)^t C_{s_t}, \tag{C.18}$$
$$y(t) = (1+g)^t y_{s_t}, \tag{C.19}$$
$$w(t) = (1+g)^t w_{s_t}, \tag{C.20}$$
$$\pi(z,t) = (1+g)^t \pi_{s_t}(z), \tag{C.21}$$
$$c(z,t) = (1+g)^t c_{s_t}(z), \tag{C.22}$$

where variables on the right hand side are growth-normalized, time invariant, variables which only depend on the state, s_t. We want to emphasize that this formulation does not impose any restriction on *off-equilibrium path* behavior. Thus, the equilibria that we exhibit also exist in the broader class.

In such an economy we immediately obtain that markups are time-invariant

$$M(z,t) = M_{s_t}(z). \tag{C.23}$$

It follows from a standard transformation, using the utility representation (equation 2), that growth-normalized variables can be determined by solving the model for a non-growing economy with a growth-adjusted personal discount rate, i.e., with

$$\hat{\delta} \stackrel{\text{def}}{=} (1+g)^{1-\gamma}\delta. \tag{C.24}$$

Intuitively, the representative agent's trade-off between consumption in different times and states is affected in identical ways by changes in the growth rate and the subjective discount factor. Thus, the effective discount rate in a growing economy, $\hat{\delta}$, depends on long term growth rates. The importance of long-term growth rates for asset pricing was recently discussed in Parlour et al. [40]. In that paper, long-term growth rates are important because they determine how much investors care about rare disaster events in the far future.

References

[1] Abreu, D., 1988, On the theory of infinitely repeated games with discounting, *Econometrica* 56, 383–96.

[2] Acemoglu, D., V. M. Carvalho, A. Ozdaglar, and A. Tahbaz-Salehi, 2012, The network origins of aggregate fluctuations, *Econometrica* 80, 1977–2016.

[3] Aghion, P., N. Bloom, R. Blundell, R. Griffith, and P. Howitt, 2005, Competition and innovation: An inverted-u relationship, *The Quarterly Journal of Economics* 120, 701–728.

[4] Bagwell, K., and R. Staiger, 1997, Collusion over the business cycle, *RAND Journal of Economics* 28, 82–106.

[5] Benassy, J.-P., 1988, The objective demand curve in general equilibrium with price makers, *The Economic Journal* 98, 37–49.

[6] Bilbiie, F. O., F. Ghironi, and M. J. Melitz, 2008, Monopoly power and endogenous product variety: Distortions and remedies, NBER Working Papers 14383, National Bureau of Economic Research.

[7] Bilbiie, F. O., F. Ghironi, and M. J. Melitz, 2012, Endogenous entry, product variety, and business cycles, *Journal of Political Economy* 120, 304 – 345.

[8] Bils, M., P. J. Klenow, and B. A. Malin, 2012, Testing for Keynesian labor demand, Nber macroeconomics annual, National Bureau of Economic Research.

[9] Calvo, G. A., 1983, Staggered prices in a utility-maximizing framework, *Journal of Monetary Economics* 12, 383–398.

[10] Campbell, J. Y., 2003, Consumption-based asset pricing, in G. M. Constantinides, M. Harris, and R. M. Stulz, eds., *Handbook of the Economics of Finance*, 805–887 (Elsevier).

[11] Chevalier, J. A., and D. S. Scharfstein, 1995, Liquidity constraints and the cyclical behavior of markups, *American Economic Review* 85, 390–96.

[12] Chevalier, J. A., and D. S. Scharfstein, 1996, Capital-market imperfections and countercyclical markups: Theory and evidence, *American Economic Review* 86, 703–25.

[13] Christiano, L. J., M. Eichenbaum, and C. L. Evans, 2005, Nominal Rigidities and the Dynamic Effects of a Shock to Monetary Policy, *Journal of Political Economy* 113, 1–45.

[14] dal Bo, P., 2007, Tacit collusion under interest rate fluctuations, *The RAND Journal of Economics* 38, 533–540.

[15] De Loecker, J., 2011, Recovering markups from production data, *International Journal of Industrial Organization* 29, 350–355.

[16] Dhingra, S., and J. Morrow, 2012, The impact of integration on productivity and welfare distortions under monopolistic competition, CEP Discussion Papers dp1130, Centre for Economic Performance, LSE.

[17] Dixit, A. K., and J. E. Stiglitz, 1977, Monopolistic competition and optimum product diversity, *American Economic Review* 67, 297–308.

[18] Duffie, D., 2001, *Dynamic Asset Pricing Theory* (Princeton University Press, Princeton, 3rd. edn.).

[19] Edmond, C., V. Midrigan, and D. Y. Xu, 2012, Competition, markups, and the gains from international trade, Working Paper 18041, National Bureau of Economic Research.

[20] Epifani, P., and G. Gancia, 2011, Trade, markup heterogeneity and misallocations, *Journal of International Economics* 83, 1–13.

[21] Fernandez-Villaverde, J., P. Guerron-Quintana, K. Kuester, and J. Rubio-Ramirez, 2011, Fiscal Volatility Shocks and Economic Activity, PIER Working Paper Archive 11-022, Penn Institute for Economic Research, Department of Economics, University of Pennsylvania.

[22] Gabaix, X., 2011, The granular origins of aggregate fluctuations, *Econometrica* 79, 733–772.

[23] Gabszewicz, J. J., and J. P. Vial, 1972, Oligopoly 'a la cournot' in general equilibrium analysis, *Journal of Economic Theory* 4, 381–400.

[24] Gali, J., 1994, Monopolistic competition, business cycles, and the composition of aggregate demand, *Journal of Economic Theory* 63, 73–96.

[25] Goodfriend, M., and R. G. King, 1998, The new neoclassical synthesis and the role of monetary policy, Working Paper 98-05, Federal Reserve Bank of Richmond.

[26] Haltiwanger, J., and J. Harrington, Joseph E., 1991, The impact of cyclical demand movements on collusive behavior, *The RAND Journal of Economics* 22, 89–106.

[27] Holmes, T. J., W.-T. Hsu, and S. Lee, 2013, Allocative efficiency, mark-ups, and the welfare gains from trade, Working Paper 19273, National Bureau of Economic Research.

[28] Hsieh, C.-T., and P. J. Klenow, 2009, Misallocation and manufacturing tfp in china and india, *The Quarterly Journal of Economics* 124, 1403–1448.

[29] Jaimovich, N., 2007, Firm dynamics and markup variations: Implications for sunspot equilibria and endogenous economic fluctuations, *Journal of Economic Theory* 137, 300–325.

[30] Jaimovich, N., and M. Floetotto, 2008, Firm dynamics, markup variations and the business cycle, *Journal of Monetary Economics* 55, 1238–1252.

[31] Jovanovic, B., 1987, Micro shocks and aggregate risk, *Quarterly Journal of Economics* 102, 395–409.

[32] Kung, H., and L. Schmid, 2014, Innovation, growth, and asset prices, *Journal of Finance* Forthcoming.

[33] Kydland, F. E., and E. C. Prescott, 1982, Time to build and aggregate fluctuations, *Econometrica* 50, 1345–1370.

[34] Lerner, A. P., 1934, The concept of monopoly and the measurement of monopoly power, *The Review of Economic Studies* 1, 157–175.

[35] Long, B., and C. I. Plosser, 1983, Real business cycles, *Journal of Political Economy* 91, 39–69.

[36] Marschak, T., and R. Selten, 1974, *General Equilibrium with price-making firms, Lecture Notes in Economics and Mathematical Systems* (Springer-Verlag, Berlin).

[37] Mehra, R., and E. C. Prescott, 1985, The equity premium: A puzzle, *Journal of Monetary Economics* 15, 145–161.

[38] Nekarda, C. J., and V. A. Ramey, 2013, The cyclical behavior of the price-cost markup, NBER Working Paper 19099, National Bureau of Economic Research.

[39] Opp, M. M., 2010, Tariff wars in the ricardian model with a continuum of goods, *Journal of International Economics* 80, 212–225.

[40] Parlour, C., R. Stanton, and J. Walden, 2011, Revisiting asset pricing puzzles in an exchange economy, *Review of Financial Studies* 24, 629–674.

[41] Peters, M., 2013, Heterogeneous mark-ups, growth and endogenous misallocation, Working paper, London School of Economics.

[42] Ravn, M., S. Schmitt-Grohe, and M. Uribe, 2006, Deep habits, *Review of Economic Studies* 73, 195–218.

[43] Rotemberg, J. J., and G. Saloner, 1986, A supergame-theoretic model of price wars during booms, *American Economic Review* 76, 390–407.

[44] Rotemberg, J. J., and M. Woodford, 1992, Oligopolistic pricing and the effects of aggregate demand on economic activity, *Journal of Political Economy* 100, 1153–1207.

[45] Samuelson, P., 1949, *Foundations of Economic Analysis* (Harvard University Press, Cambridge).

[46] Schmitt-Grohe, S., 1997, Comparing four models of aggregate fluctuations due to self-fulfilling expectations, *Journal of Economic Theory* 72, 96–147.

[47] van Binsbergen, J. H., 2014, Good-specific habit formation and the cross-section of expected returns, *Journal of Finance* Forthcoming.

[48] Woodford, M., 1986, Stationary sunspot equilibria: The case of small fluctuations around a deterministic steady state, Technical report, University of Chicago, Unpublished manuscript.

[49] Woodford, M., 2003, *Interest and Prices: Foundations of A Theory of Monetary Policy* (Princeton University Press, Princeton, N.J.).

[50] Woodford, W., 1991, Self-fulfilling expectations and fluctuations in aggregate demand, in N. Mankiw, and D. Romer, eds., *New Keynesian Economics* (MIT Press, Cambridge, MA).

[51] Zhelobodko, E., S. Kokovin, M. Parenti, and J.-F. Thisse, 2012, Monopolistic competition: Beyond the constant elasticity of substitution, *Econometrica* 80, 2765–2784.