Endogenous Firm Competition and the Cyclicality of Markups[*]

Hassan Afrouzi[†] Luigi Caloi[‡]

November 8, 2022

Abstract

We show that the cyclicality of *output growth* is a sufficient predictor for the cyclicality of markups in models that micro-found variable markups through dynamic trade-offs. We use data on markups from the U.S. as well as survey data on firms' expectations from New Zealand to test the predictions of these models and find evidence in favor of their mechanisms. Finally, we study the implications of these mechanisms for cyclicality of markups in a calibrated general equilibrium model. In particular, we find that the degree of hump-shaped response in output is crucial for the direction of aggregate markup cyclicality.

Keywords: markup cyclicality, implicit collusion models, customer-base models

JEL Classification: E3

[*] We are grateful to the editor and three anonymous referees for their thoughtful comments and suggestions. We also thank Olivier Coibion, Saroj Bhattarai, Andrew Glover, Matthias Kehrig, as well as seminar participants at UT Austin, Midwest Macro Meeting, and Federal Reserve Banks of Dallas and Kansas City.

[†] Columbia University, Department of Economics, 420 West 118 St. New York, NY 10027. Email: hassan.afrouzi@columbia.edu.

[‡] Columbia University, Department of Economics, 420 West 118 St. New York, NY 10027.

1 Introduction

The cyclicality of markups is a key transmission mechanism in the business cycle literature. Countercyclical markups, for instance, are broadly viewed as a propagation mechanism in New Keynesian models (Christiano, Eichenbaum and Rebelo, 2011), as well as a potential reason for positive comovement between hours and wages (Rotemberg and Woodford, 1992). Nonetheless, theories that micro-found variable markups have conflicting predictions regarding their cyclical behavior.[1]

In this paper, we provide a unified law of motion for markups in a broad class of models that micro-found variable markups through *dynamic trade-offs*—namely implicit collusion and customer-base models—and show that cyclicality of *output growth* is a sufficient predictor for the cyclicality of markups in these models.

We formally show that, up to a first-order approximation, both implicit collusion and customer-base models yield the same reduced form expression for the dynamics of markups. Specifically, at the firm level, current markups depend on the net present value of all expected sales growths in the future. Once aggregated, the net present value of future sales collapses to the net present value of output growth in the economy. Therefore, the two models relate current markups to expected output growths in the future. They differ, however, in terms of the sign restrictions that they imply for this reduced-form representation of markup dynamics. Hence, one can test and differentiate between these two sets of models by estimating the common law of motion for markups and comparing the empirical estimates with the sign restrictions implied by each model.

Our second contribution is to use data from Compustat to test these predictions for the U.S. To do so, we estimate firm-level markups following the methodology of De Loecker et al. (2020) and subsequently estimate the law of motion for markups. First, we find markups comove *positively* with

[1] In particular, two common micro-foundations—implicit collusion and customer-base models—yield different cyclicalities: implicit collusion models are interpreted as implying countercyclical markups, while customer-base models have been used to generate both procyclical and countercyclical markups.

firms' future sales (using lagged values of sales and markups as instruments for firms' expectations of future sales growth and markups), which provides support for the implicit collusion model.

Second, to further test the implications of these two models, we also study the heterogeneity of these results in the cross-section of firms. Both models imply that the magnitude of the relationship between markups and future sales growths should be decreasing in firms' number of competitors. A second prediction of the two models is that the relationship between markups and expected future sales growths should be stronger for firms that value future profits more. Using insights from previous work on financial frictions (e.g. Gilchrist et al., 2017), we hypothesize that firms with a higher debt-to-asset ratio should value future profits less relative to current profits due to a higher preference for liquidity in short-run, higher effective interest rates or higher anticipated exit rates.[2] Therefore, according to the models, all else equal, we should see a weaker relationship between markups and expected future sales growths among more leveraged firms. We find both predictions consistent with the data: the coefficient on future sales growth for firms on the upper percentiles of lag log markup and negative of lag of debt-to-asset ratio is positive and significantly larger than the estimate using all other observations, and more than thrice the size of the coefficient reported in our benchmark regression.

Finally, to examine the external validity of these findings outside of the Compustat, we estimate the law of motion for markups using survey data on firms' expectations from New Zealand, introduced by Coibion et al. (2018). This approach also allows us to directly rely on data on the expected sales growth of firms rather than the actual realization of sales. Consistent with the findings in the Compustat sample, the results favor the implicit collusion model. In particular, we find a positive and significant relationship between markups and expectations of future sales among firms with fewer competitors.

Therefore, both sets of empirical evidence are consistent with the predictions of implicit collusion models but at odds with customer-base models.

Our final contribution is to calibrate a general equilibrium model with implicit collusion and

[2] We thank an anonymous referee for highlighting the default mechanism.

study the implications of these findings for the cyclicality of markups. We find that depending on how hump-shaped the response of output is, markups can either be procyclical or countercyclical in these models. The reason is that the implicit collusion model relates markup cyclicality to output growth rather than output itself. Hence, if output growth is expected to be positive in an expansion (due to a hump-shaped response), then markups increase on impact and covary positively with output. However, if there is no hump shape in output response, then after an expansionary shock where output jumps to a high level and is expected to fall back to its steady state after that, output growth is expected to be negative, and markups decrease in expansions.

Thus, accounting for the hump-shaped response of output to shocks, as observed in the empirical literature,[3] is instrumental for the cyclicality of markups. We illustrate the importance of this mechanism by examining the cyclicality of markups conditional on TFP shocks in the model and relating it to the empirical estimates in Nekarda and Ramey (2020), who find that markups are procyclical conditional on TFP shocks. First, we show that a model without a hump-shaped response of output delivers the wrong cyclicality: when productivity goes up, and output rises on impact, output growth is expected to be negative, and markups fall, leading to countercyclical markups conditional on TFP shocks. However, once we allow for a hump-shaped response for output (using investment adjustment costs), we find that markups become procyclical conditional on TFP shocks and match the conditional correlation of markups and output relatively well as non-targeted moments.

Literature Review. Both implicit collusion and customer-base models are used within macroeconomic models to study the markup setting behavior of firms. In our analysis, we start by building the firm side of the implicit collusion model, and show that markups are determined by the joint distribution of expected growth of output and stochastic discount rates. This allows us to reconcile the seemingly contradictory predictions of these models in a unified framework. For instance,

[3] See, e.g., Ramey (2011); Ramey and Shapiro (1998); Monacelli and Perotti (2008) for government spending shocks, Sims (2011); Smets and Wouters (2007) for productivity shocks, and Christiano et al. (2005) for monetary policy shocks.

Rotemberg and Saloner (1986) assume that demand shocks are i.i.d., implicitly implying that the expected demand growth is countercyclical, and conclude that markups are countercyclical. On the other hand, Kandori (1991); Haltiwanger and Harrington Jr (1991); Bagwell and Staiger (1997), each by assuming alternative processes for demand shocks find that these models can produce procyclical markups.

Rotemberg and Woodford (1991, 1992) are the first to study the implicit collusion model within a DSGE model. Contrary to the partial equilibrium models, their general equilibrium setting endogenously pins down the joint distribution of output growth and stochastic discount rates, which gives rise to countercyclical markups; however, their result is not robust to the structure of the shocks and is reversed by introducing a hump-shaped response for output.

Phelps and Winter (1970) is the first paper that formalizes the idea for customer-base models. Various papers have used this idea to study the cyclical behavior of markups. Different versions of customer-base models have been shown to create either procyclical or countercyclical markups. For instance, by micro-founding the game between firms and customers, Paciello et al. (2018) find that markups are procyclical, but Ravn et al. (2006) argue that they are countercyclical. In this paper, we do not take a stance on the micro-foundations of this friction.[4] Instead, using a simple customer-base model with an exogenous habit formation process on the side of customers, we show that markups can be either pro- or countercyclical depending on whether the response of output to shocks is hump-shaped or not.

Another class of models that generate variable markups use demand systems where the elasticity of demand varies with firms' size, either by directly assuming Marshall's second law of demand using the aggregator function as in Kimball (1995), or by assuming nested CES aggregators, as in Atkeson and Burstein (2008). A recent application of Kimball preferences is Edmond et al. (2018) who use this demand system to assess the cost of markups, whereas a recent application of the latter

[4] There has been a tremendous amount of progress in recent years in micro-founding this friction using search and matching frameworks. See, e.g., Gourio and Rudanko (2014); Kaplan and Menzio (2016); Bornstein (2018).

is Burstein et al. (2020) who find that the cyclicality of markups depends on the level of aggregation. Moreover, New Keynesian models also have predictions for the cyclicality of markups due to price stickiness. We discuss the relationship between our paper and these sets of models in Section 5.

Outline. Section 2 introduces the firm side of implicit collusion and customer-base models and derives the unified law of motion for markups. Section 3 presents the evidence for this law of motion using Compustat data in the U.S. and survey data on firms' expectations from New Zealand. Section 4 discusses the implications of the law of motion for markups in a calibrated general equilibrium model. Section 5 discusses the strengths and weaknesses of Compustat for our exercise, potential issues with the estimation of markups as well as the relationship between our framework and other models of variable markups. Section 6 concludes. Proofs and additional tables and figures are included in a companion Online Appendix.

2 A Unified Law of Motion for Markups

In this section, we revisit customer-base and implicit collusion model, both of which micro-found markups through dynamic trade-offs. Our results in this section are twofold. First, we show that both classes of models imply fundamentally similar laws of motion for markups, where markups depend on the net present value of future *sales growth* for firms. Second, we find that despite this similarity, each model associates a different sign to this relationship: in implicit collusion models markups move positively with future sales, but in customer-base models this relationship is negative. In the remainder of this section, we introduce the firm side of implicit collusion and customer-base models respectively, and derive a unified law of motion for markups in both.

2.1 Implicit Collusion Models

We follow Rotemberg and Woodford (1991, 1992) in setting up the implicit collusion model but we depart from their representation by deriving the law of motion for markups, and showing that markups depend on the net present value of future *sales growth*.

There is a final good of consumption in the economy which is produced using a large number of intermediate differentiated goods. There is a unit measure of intermediate good sectors indexed by

$i \in [0,1]$. In each sector, there are N identical firms producing differentiated goods, indexed by i, j where $j \in \{1, \dots, N\}$.

The Final Good Producer. The final good producer takes the price of consumption good, P_t, as given and produces with

$$Y_t = \left[\int_0^1 Y_{i,t}{}^{\frac{\sigma-1}{\sigma}} di\right]^{\frac{\sigma}{\sigma-1}}, \quad Y_{i,t} = \Phi(Y_{,1}, \dots, Y_{i,N}) \equiv \left[N^{-\frac{1}{\eta}} \sum_{j=1}^{N} Y_{i,j,t}^{\frac{\eta-1}{\eta}}\right]^{\frac{\eta}{\eta-1}} \tag{1}$$

Therefore, σ is the elasticity of substitution across sectors, and η is the elasticity of substitution within sectors. The profit maximization problem of this firm leads to the following standard demand functions implied by nested CES preferences:

$$\frac{Y_{i,j,t}}{Y_t} = D(\frac{P_{i,j,t}}{P_t}; \frac{P_{i,-j,t}}{P_t}) \equiv \frac{1}{N}\left(\frac{P_{i,j,t}}{P_t}\right)^{-\eta}\left[\frac{1}{N}\sum_{k=1}^{N}\left(\frac{P_{i,k,t}}{P_t}\right)^{1-\eta}\right]^{\frac{\eta-\sigma}{1-\eta}} \tag{2}$$

where $P_t \equiv \left[\int_0^1 [N^{-1}\sum_{k=1}^{N} P_{i,k,t}^{1-\eta}]^{\frac{1-\sigma}{1-\eta}} di\right]^{\frac{1}{1-\sigma}}$ is the aggregate price level. This demand system leads to the following expression for the firm's elasticity of demand, where this elasticity is an average of σ and η weighted by the firm's market share (Atkeson and Burstein, 2008):

$$-\frac{\partial \log(Y_{i,j,t})}{\partial \log(P_{i,j,t})} = (1 - s_{i,j,t})\eta + s_{i,j,t}\sigma, \quad s_{i,j,t} \equiv \frac{P_{i,j,t} Y_{i,j,t}}{\sum_{k=1}^{N} P_{i,k,t} Y_{i,k,t}} \tag{3}$$

A general assumption in two layer CES models is that $\eta > \sigma > 1$. This assumption corresponds to the case that goods within sectors are closer substitutes than goods across sectors, and ensures that Marshall's second law of demand holds in relative terms (that demand is less elastic at higher relative quantity, or firms with higher market share have less elastic demand).

Intermediate Goods. We assume all N firms within each sector i are identical in their production technology and use capital and labor to produce with a Cobb-Douglas production function, $Y_{i,j,t} = Z_t^a K_{i,j,t}^{\alpha} L_{i,j,t}^{1-\alpha}$, where Z_t^a is an economy-wide technology shock with an AR1 process:

$$\log(Z_t^a) = \rho_a \log(Z_{t-1}^a) + \sigma_a \varepsilon_{a,t}, \quad \varepsilon_{a,t} \sim \mathcal{N}(0,1)$$

Moreover, we assume factor markets are competitive and firms take the wages, W_t, as well as the rental rate of capital R_t as given.

The Repeated Game of Sector i. Let $Y_{i,t} \equiv \Phi(Y_{i,1,t}, \ldots, Y_{i,N,t})$ denote the output of sector i at time t. Firms in any sector i take their demand functions in Equation (2) as given and discount future profits with a stochastic discount factor process, $\{\beta^t Q_{0,t} : t \geq 0\}$,[5] and play the following infinitely repeated game.

As in every super-game, this repeated game has many potential equilibria. Although there is no rigorous way to rank the multiple equilibria of this game, the literature on the implicit collusion models focuses on the equilibrium in which firms earn the highest possible profit stream subject to an incentive compatibility constraint that eliminates every firm's incentive to deviate from the equilibrium strategy of collusion.[6] Following this literature, we also assume that there exists a perfect monitoring system that detects any deviations with probability one. Therefore, the best cheating strategy for a firm is to best respond to collusion outputs of their rivals, knowing that it will trigger the punishment sub-game.[7]

Characterization of the Repeated Game Equilibrium. The equilibrium strategy is constructed as follows: at time 0 firms form the following contingent plan for all possible states in the future. For every single state at every point of time, every firm chooses a markup that yields the highest profit for the sector and is incentive compatible with collusion relative to the following punishment

[5] In general equilibrium, this process is determined by the relative marginal utilities of households in different states.

[6] Nonetheless, such an equilibrium is not necessarily the equilibrium with the highest net present value of profits; it may be the case that occasional deviations yield higher profits compared to staying in collusion forever. Therefore, there might be an equilibrium with occasional collusion that dominates the best equilibrium in which firms always collude. We abstract from this case, following Rotemberg and Woodford (1991, 1992, 1999).

[7] In the absence of such a system, however, static best responding may not be the best cheating strategy for a firm. If small deviations were unnoticeable with some probability, characterizing the best strategy is nontrivial. For instance, in an environment with imperfect monitoring, Green and Porter (1984) characterize equilibria in which firms switch to punishment when their price falls below a trigger price, even if it is due to a negative demand shock rather than a cheating competitor.

strategy: in case a firm deviates from the agreement, the game enters a punishment stage where all firms will charge the static best response markup forever after. However, at each period there is a possibility that the industry will renegotiate this with probability $1 - \gamma$ and will move back to the collusion stage. This probability γ pins down the expected punishment duration. The industry expects to remain in the punishment stage for an average of $1/(1 - \gamma)$ periods.

Therefore, firms within every sector maximize the discounted value of the industry's lifetime profits such that no firm in no state has the incentive to cheat.[8] Note that incentive compatibility is the only restricting concern in this setting. Without it, firms would choose the monopoly markup for the industry in every state. However, a firm's incentive to cheat is at its highest level when the rest of the firms are committed to producing the monopoly output of the industry. This incentive declines as the markups of the other firms decrease towards the one in the static best-response equilibrium. Moreover, since firms choose their markups to be incentive compatible with collusion in every possible state, the game stays in the collusion sub-game forever from which no one has the incentive to deviate. Therefore, the proposed strategy is a sub-game perfect Nash equilibrium.

Moreover, with a CRS production function, firms' capital-to-labor ratios are independent of their output level. This can be interpreted as firms having a constant marginal cost of production in a given period that is pinned down by factor prices and is independent of firms' level of production, or their total demand:

$$MC_t = \frac{1}{Z_t^a}\left(\frac{R_t}{\alpha}\right)^{\alpha}\left(\frac{W_t}{1-\alpha}\right)^{1-\alpha} \tag{4}$$

Therefore, given firms' price-taking behavior in factor markets, what pins down a firm's size and level of production is their market power: given their demand structure, a choice for prices determines firms' demand, which then implies a certain level of production to meet that demand.

Since we only focus on symmetric equilibria, in characterizing the strategy of a firm, we only consider strategies in which a firm's competitors all are charging the same markup which is going

[8] Note that this does not require any explicit collusion among firms as each firm simply chooses the highest markup from which no one in the industry has the incentive to deviate.

to be the collusion markup in the equilibrium. Therefore, firm i, j's profit from the action profile $\mu_i^t \equiv (\mu_{i,j,t}; \mu_{i,t})$, where $\mu_{i,t}$ is the collusion markup chosen by the industry and $\mu_t \equiv P_t/MC_t$ is the average markup in the economy, is given by

$$\Pi_{i,j,t}(\mu_{i,j,t}; \mu_{i,t}) = P_t Y_t (\frac{\mu_{i,j,t} - 1}{\mu_{i,t}}) \left(\frac{\mu_{i,t}}{\mu_t}\right)^{1-\sigma} D(\frac{\mu_{i,j,t}}{\mu_{i,t}}; 1) \tag{5}$$

Since the model is real and all sectors are symmetric, henceforth we normalize aggregate price to one, $P_t \equiv 1$. The following Proposition formalizes the equilibrium.

Proposition 1. *Each firm in sector i, maximizes its net present value of future profits subject to no other firm having the incentive to undercut them:*

$$\begin{aligned} &\max_{\{\mu_{i,t}\}_{t=0}^{\infty}} \tfrac{1}{N}\mathbb{E}_0 \textstyle\sum_{t=0}^{\infty} (\beta\gamma)^t Q_{0,t} Y_t (1 - \tfrac{1}{\mu_{i,t}}) \mu_{i,t}^{1-\sigma} \mu_t^{\sigma-1} \\ s.t. \quad &\max_{\rho_{i,t}} \left\{(\rho_{i,t} - \tfrac{1}{\mu_{i,t}}) D(\rho_{i,t}; 1)\right\} - \tfrac{1}{N}(1 - \tfrac{1}{\mu_{i,t}}) \leq \beta\gamma \mathbb{E}_t Q_{t,t+1} \tfrac{Y_{t+1}}{Y_t} \left(\tfrac{\mu_{t+1}/\mu_{i,t+1}}{\mu_t/\mu_{i,t}}\right)^{\sigma-1} \Gamma_{i,t+1} \quad (6) \\ &\Gamma_{i,t} \equiv \tfrac{1}{N}\left[(1 - \tfrac{1}{\mu_{i,t}}) - \mu_s^{-\sigma}(\mu_s - 1)\mu_t^{\sigma-1}\right] + \beta\gamma \mathbb{E}_t Q_{t,t+1} \tfrac{Y_{t+1}}{Y_t} \left(\tfrac{\mu_{t+1}/\mu_{i,t+1}}{\mu_t/\mu_{i,t}}\right)^{\sigma-1} \Gamma_{i,t+1} \end{aligned}$$

where $\beta^\tau Q_{t,t+\tau}$ is the time t price of a claim that pays a unit of consumption at $t + \tau$, and $\mu_s \equiv \frac{(N-1)\eta+\sigma}{(N-1)\eta+\sigma-N}$ is the equilibrium markup of static best responding for firms at any state. $\eta > \sigma$ guarantees that $\frac{\eta}{\eta-1} \leq \mu_s \leq \frac{\sigma}{\sigma-1}$. The solution to this problem $\{\mu_{i,t}\}_{t=0}^{\infty}$ exists, and it is a Sub-game Perfect Nash Equilibrium for the repeated game in sector i, in which firms always collude.

Equation (6) is the incentive compatibility constraint which requires that all firms in a sector prefer collusion to cheating in every possible state. Accordingly, such a sequence of assigned collusion markups are incentive compatible by construction and therefore form an equilibrium.

Now, suppose that the model is calibrated such that the constraint binds in the steady state (otherwise, the oligopoly acts as a monopoly and the model essentially becomes a monopolistic competition model across sectors). Then for small perturbations around that steady state, a first-order approximation yields

$$\hat{\mu}_t = \psi_1 \mathbb{E}_t \left[\Delta \hat{y}_{t+1} + \hat{q}_{t,t+1}\right] + \psi_2 \mathbb{E}_t \left[\hat{\mu}_{t+1}\right] \tag{7}$$

where $\Delta \hat{y}_{t+1} \equiv \frac{\Delta Y_{t+1}}{Y}$ is percentage growth of sales with respect to the steady state output, and

$\hat{q}_{t,t+1} \equiv \frac{Q_{t,t+1}-\bar{Q}}{\bar{Q}}$ is percentage deviation of the stochastic discount rate from its steady state level. Moreover,

$$\psi_1 \equiv \gamma\beta \frac{\bar{\mu}\bar{\Gamma}}{D(\bar{\rho};1)-1/N} \geq 0 \tag{8}$$

$$\psi_2 \equiv \gamma\beta \frac{D(\bar{\rho};1)-(\sigma-1)(\mu_C-1)(\frac{\bar{\mu}}{\mu_C})^{\sigma}/N}{D(\bar{\rho};1)-1/N} \lesseqgtr 0 \tag{9}$$

Equation (7) gives the law of motion for average markups in the partial equilibrium of the firm side in this economy. This is the key equation in this paper that will underlie all the results in later sections. Therefore, the following subsection is devoted to discussing this result.

Interpretation. Implicit collusion implies that markups are forward-looking variables that depend on the expected change in demand in the next period, the changes in the price of future profits, and the expected change of markup in the next period. ψ_1, which is the coefficient on the first two, is a positive number that is increasing in steady state gains from collusion ($\gamma\beta\bar{\mu}\bar{\Gamma}$) and decreasing in the marginal revenue that a firm makes by cheating in the steady state ($D(\bar{\rho};1)-1/N$). The intuition behind this equation is the key to understanding the main results of this paper. Two things between current period and the period ahead affect the current period's markup: first, the current price of next period's profit, which is the discount factor of the firms. The more patient the firms are in an industry, the higher their collusion markup will be today as they value future profits more. Second, the expected growth in demand from current period to next period. If firms expect that demand tomorrow will be higher than today, then they do not want to lose the chance of cheating tomorrow by cheating today. Basically, firms want to wait until demand is at its highest to take advantage of cheating, as in that case they will collect the highest cheating gains. This incentive to wait diminishes firms' cheating incentives in the current period, allowing the industry to sustain a higher collusion markup. Therefore, when firms expect output to grow, they will charge markups that are closer to the monopoly one.

ψ_2, however, can theoretically be positive or negative based on the calibration of the model. The reason is that there are two opposite forces that affect the firms' cheating incentives based on their expectation of future markup. Suppose firms expect that the markup will be higher in the future than its steady-state level. On one hand, they do not want to miss the chance of charging such high markups in their own industry by cheating today and entering the punishment subgame. On the other hand, since all other industries will also charge high markups, firms expect a large demand shift towards their industry in the punishment sub-game where their markup is lower, which gives an incentive firms to push the industry to the punishment subgame by cheating today. The magnitude of this effect depends on the elasticity of substitution across industries; as seen in the expression of ψ_2, when σ is close to 1, this force is negligible.

The previous results in this literature can be seen as special cases of Equation 7. For example, Rotemberg and Saloner (1986) setup can be seen as the case where $\hat{q}_{t,t+1} = 0$ due to a constant discount rate, and $\mathbb{E}_t \left[\Delta \hat{y}_{t+1}\right] = -\hat{y}_t$ as shocks are assumed to be i.i.d. over time. Therefore, in their model, $\hat{\mu}_t = -\psi_1 \hat{y}_t$, which is a demonstration of their result that markups should be counter-cyclical. But as (7) implies, assuming other processes for these variables can give rise to different results. With two different random processes, $\Delta \hat{y}_{t+1}$ and $\hat{q}_{t,t+1}$, that are potentially correlated, the spectrum of possibilities for their underlying distribution is large enough to allow for *any* type of result in terms of the cyclicality of markups. Therefore, we need to pin down this joint distribution, which in the case of this paper will be done by introducing a household side for the model.

2.2 Customer-base Models

Another class of models that micro-found variable markups is based on the notion that there is inertia in how fast customers shift their demand across firms (Phelps and Winter, 1970).[9] Accordingly, firms' pricing decisions affect their market share in the future. The dynamics of markups in these models depend on how customers are reacting to the pricing of the firms over time.

In this section, we build a simple reduced-form customer-base model where the inertia in demand comes from habit formation on the customer side. We then show that these models imply a similar law of motion for markups as the implicit collusion models but have different predictions for the sign of coefficients on output growth and stochastic discount rates.

[9] See Bornstein (2018) for a micro-foundation with an application to recent trends in business dynamism and markups.

Model Specification. Consider the final good producer of Section 2.1. To incorporate the customer-base model, we assume that this final good producer forms external habits over the goods within industries, meaning

$$Y_t = \left[\int_0^1 Y_{i,t}{}^{\frac{\sigma-1}{\sigma}} di\right]^{\frac{\sigma}{\sigma-1}}, \quad Y_{i,t} \equiv \left[N^{-\frac{1}{\eta}} \sum_{j=1}^{N} S_{i,j,t}^{\frac{1}{\eta}} Y_{i,j,t}^{\frac{\eta-1}{\eta}}\right]^{\frac{\eta}{\eta-1}}$$

Where $S_{i,j,t}$ is the external habit of producer in using the input $Y_{i,j,t}$, which is taken as given by the final good producer at time t. We assume that $S_{i,j,t}$ has the following general law of motion

$$S_{i,j,t} = \gamma h(\frac{\mu_{i,j,t}}{\mu_{i,t}}) S_{i,j,t-1} + 1 - \gamma$$

where $h(.)$ is differentiable, $h(1) = 1$, $h'(.) < 0$, and $\gamma \in [0, 1)$. Notice that $\gamma h(\frac{\mu_{i,j,t}}{\mu_{i,t}})$ is how fast the habit of the final good producer depreciates over time, $h'(.) < 0$ implying that this depreciation is faster if the firm charges a higher markup relative to its competitors (Phelps and Winter (1970) interpret $S_{i,j,t}$ as the measure of customers matched to the firm at time t, in which case $\gamma h(\frac{\mu_{i,j,t}}{\mu_{i,t}})$ is the separation rate of the firm's customers, which increases with the firm's markup).[10]

Solving the final good producer's problem now implies the following demand structure:

$$Y_{i,j,t} = Y_t S_{i,j,t} D(P_{i,j,t}; P_{i,-j,t})$$

where $D(.;.)$ is defined exactly as in Section 2.1. Firm i, j takes demand as given and maximizes the net present value of all its future profits by choosing a relative markup $\frac{\mu_{i,j,t}}{\mu_{i,t}}$, and $S_{i,j,t}$, where $S_{i,t}$ is the final good producer's habit for the others in the sector, in the symmetric equilibrium. Therefore, firm i, j's dynamic problem is

$$\begin{aligned} &\max_{\{\mu_{i,j,t}, S_{i,j,t}\}_{t=0}^{\infty}} \mathbb{E}_0 \textstyle\sum_{t=0}^{\infty} \beta^t Q_{0,t} Y_t S_{i,j,t} \left(\frac{\mu_{i,t}}{\mu_t}\right)^{1-\sigma} (\frac{\mu_{i,j,t}-1}{\mu_{i,t}}) D(\frac{\mu_{i,j,t}}{\mu_{i,t}}; 1) \\ s.t. \quad & S_{i,j,t} = \gamma h(\tfrac{\mu_{i,j,t}}{\mu_{i,t}}) S_{i,j,t-1} + 1 - \gamma \end{aligned}$$

Proposition 2. *In a symmetric equilibrium where all firms identically solve the problem above, the law of motion for markups, up to a first order approximation, takes the same form as the implicit*

[10] For a more recent model of customers as demand shifters see Afrouzi et al. (2020).

collusion model, i.e.

$$\hat{\mu}_t = \psi_1 \mathbb{E}_t[\hat{q}_{t,t+1} + \Delta \hat{y}_{t+1}] + \psi_2 \mathbb{E}_t[\hat{\mu}_{t+1}] \tag{10}$$

where

$$\psi_1 \equiv -\frac{\beta\gamma}{1+\zeta}\frac{\zeta(1-\mu_s^{-1})}{\mu_s^{-1}(1-\beta\gamma)+\zeta} \leq 0 \tag{11}$$

$$\psi_2 \equiv \frac{\beta\gamma}{1+\zeta} \geq 0 \tag{12}$$

with $\zeta \equiv -\frac{\gamma h'(1)}{(1-N^{-1})\eta + N^{-1}\sigma} > 0$ *and* $\mu_s = \frac{\eta(1-N^{-1})+\sigma N^{-1}}{(\eta-1)(1-N^{-1})+(\sigma-1)N^{-1}}$ *being the markup of a firm with no inertia in their demand. Moreover, the magnitudes of* ψ_1 *and* ψ_2 *decrease with the number of firms within sectors.*

Equation (10) formalizes the idea that the law of motion implied by this model takes the same form as that of the implicit collusion model. In these models, firms' markups are variable because their incentives to invest in their customer base vary with their expectations of how much demand there will be in the future. Again, the trade-off boils down to a relative demand argument: if the firm expects demand to be larger in the future relative to today, they have a higher incentive to invest in their customer base and vice-versa. Therefore, what determines today's markup is the firms' expectation of its sales growth, which implies a similar law of motion for markups as in the implicit collusion model, but the signs of ψ's are different as shown in Equations (11) and (12).

In contrast to the implicit collusion model where markups move positively with the expected growth of sales, the customer-base model assigns a negative relationship between the two. The reason is that expected higher demand in the future induces firms to invest more in their customer bases: they reduce their markups to lure in demand, expecting that they will "harvest" a higher demand when aggregate demand is expected to be higher.[11]

[11] See Bornstein (2018) for a detailed and recent discussion of investment and harvesting motives of firms when there is inertia in demand.

3 Testing the Law of Motion for Markups

The goal of this section is to empirically test this law of motion and provide evidence on the signs and magnitudes of coefficients ψ_1 and ψ_2. To do so, we use data from Compustat as well as data on firms' expectations from New Zealand (introduced by Coibion et al. 2018) to study the relationship between firms' markups and their future (expected) sales.

3.1 Model Predictions

Before presenting our empirical evidence, we draw a series of predictions that are implied by the two models in the previous section to guide our empirical strategy. The following statements summarize these predictions:

1. Both implicit collusion and customer base models relate markups to the net present value of firms' future sales and imply a law of motion for markups of the form

   $$\hat{\mu}_t = \psi_1 \mathbb{E}_t[\Delta \hat{y}_{t+1} + \hat{q}_{t,t+1}] + \psi_2 \mathbb{E}_t[\hat{\mu}_{t+1}]$$

   where the implication of each model for the signs of the coefficients are different. The following table summarizes these implications:

	Implicit Collusion	Customer Base
ψ_1	> 0	< 0
ψ_2	$\lesseqgtr 0$	> 0

   Table 1: Sign of Coefficients in the Law of Motion for Markups in Different Models

2. In both models, the magnitude of the coefficient on future sales ($|\psi_1|$) decreases with a firm's number of competitors.

3. In both models, the magnitude of coefficients ψ_1 and ψ_2 decrease with the firms' discount factor.

3.2 Evidence from the United States

In order to analyze firms' markups in the U.S., we have exploited the richness of the firm-level balance sheet information from Compustat dataset. While this dataset has the caveat of covering only publicly listed firms, it contains detailed financial data covering a long period of time for a broad section of the economy—it contains yearly data from 1950 to 2016 and it covers 29$\%$ of private U.S. employment (Davis et al., 2006).[12] As discussed in more detail in Online Appendix E, following De Loecker et al. (2020), we estimate markups as:

$$\mu_{it} = \theta^v_{i,t} \frac{P_{i,t} Q_{i,t}}{P^v_{i,t} V_{i,t}},$$

where $V_{i,t}$ is a variable input of production, $Q_{i,t}$ is output and $\theta^v_{i,t}$ is the output elasticity of the variable input $V_{i,t}$. To estimate $\frac{P_{i,t}Q_{i,t}}{P^v_{i,t}V_{i,t}}$, we use the variables *Sales* and *Cost of Goods Sold (COGS)* from Compustat, which are further explained in Online Appendix D. In order to estimate the output elasticity of the variable input, we used the production function estimation method also used in De Loecker et al. (2020) with some small changes.[13] Table 2 provides summary statistics for firms' sales and markups in the data. [TABLE 2 HERE]

3.2.1 Benchmark Specification and Results for Prediction 1

Having measures of markups and sales growth in the data (denoting them $M_{i,t}$ and $\Delta \log(Sales_{i,t})$ for firm i at time t), we estimate the following specification:

$$\log(M_{i,t}) = \phi_1 \Delta \log(Sales_{i,t+1}) + \phi_2 \log(M_{i,t+1}) + \varepsilon_{i,t} \tag{13}$$

where we instrument for expectations of $\Delta \log(Sales_{i,t+1})$ and $\log(M_{i,t+1})$ using the following GMM condition:

$$\mathbb{E}_t[(\log(M_{i,t}) - \phi_1 \Delta \log(Sales_{i,t+1}) - \phi_2 \log(M_{i,t+1}))\mathbf{z}_{t-1}] = 0 \tag{14}$$

[12] In Section 5 and Online Appendix E we discuss the strengths and caveats of this dataset to our study.

[13] Specifically, following Traina (2018) we estimated time-invariant but industry-specific output elasticities. We used SIC 2-digit codes for the definition of industries.

Here, $\mathbf{z}_{t-1}$ includes four lags of log of sales and log of markups for the firm dated $t-1$ and before. This instrumental variable approach is necessary because the model predicts that markups are forward-looking and are determined by firms' *expectations* of their future sales growth. Since we do not observe firms' expectations, assuming rational expectations, we use the realized values of these variables as proxies for firms' forecasts of these variables at time t. Nonetheless, this creates an endogeneity problem because any shocks to markups or sales between time t and $t+1$ are orthogonal to expectations at time t but correlate with realized values of these variables, hence biasing the estimates. Using lags of sales and markups allows us to eliminate this concern by only utilizing the variation in $\Delta \log(Sales_{i,t+1})$ and $\log(\mu_{i,t+1})$ that is predictable at time t, but orthogonal to any shock to these variables after time t.[14]

The results of this estimation exercise are reported in Table 3. Column (3), which reports IV-GMM estimates with both time and SIC 2-digit industry fixed effects, constitutes our benchmark results for the U.S. and shows that markups are positively correlated with firms' future expected sales growths and future markups. For reference, Column (1) shows the OLS estimates, and Column (2) reports IV-GMM estimates with time but without SIC 2-digit industry fixed effects. It is important to note that the OLS estimates give the opposite sign, indicating the necessity of our instrumental variable approach. [TABLE 3 HERE]

The positive signs on the coefficients for future expected sales and markups are consistent with the implicit collusion model but go against the predictions of the customer-base models (Prediction 1). Having this in mind, we examine the heterogeneity of the effects based on size and discount factor next (Predictions 2 and 3).

Heterogeneity Based on Markups and Debt-to-Asset Ratio (Predictions 2 and 3). One major prediction of both models is that the magnitude of the coefficient on relative sales should decrease with a firm's number of competitors (Prediction 2). Given the direct relationship between average markups and the number of competitors in the models, we expect firms with higher average markups

[14] This is a common approach to deal with such endogeneity issues in estimating forward-looking equations in macroeconomics. See, for instance, Galí and Gertler (1999).

to have a stronger relationship between current markups and expected future sales. Thus, we use the lag of log markups as our main proxy for $1/N$ in the model. Since $1/N$ also represent firms' sales shares in the model, as an alternative measure, we also use firms' relative size as an alternative proxy for $1/N$. In Section 5, we provide further discussion about the challenges of proxying concentration or markups in the Compustat data.

Another related prediction of the two models is that markups should be more sensitive to future sales of firms if they are more patient, meaning that they assign higher values to future profits relative to contemporaneous profit (Prediction 3). While we do not observe discount rates directly, it is reasonable to assume that financially constrained firms might be more subject to default risk and assign lower relative value to future profits due to a higher cost of borrowing (see, e.g., Gilchrist et al., 2017). To the extent that higher leverage is associated with a higher cost of borrowing for firms, their future cash flow is discounted more heavily due to higher effective interest rates which cause these firms to discount the future more. Similarly, higher leverage might also increase the probability of default in the future and thus lead to further discounting due to a higher probability of exit for the firms.[15] To link the debt-to-asset ratio in Compustat to default risk, we relied on the work done by Ottonello and Winberry (2020). Using Compustat, they construct two measures of default risk, the debt-to-asset ratio and an estimated probability of default. Even after controlling for firm fixed effects and sector-by-quarter fixed effects, they show that debt-to-asset ratio is negatively correlated with other the probability of default and S&P's long-term issue rating of the firm, which gives them (and us) confidence that debt-to-asset can proxy default risk. Taking these results together, Prediction 3 would imply that we should see a smaller sensitivity of markups to future sales among more leveraged firms.

[15] In contrast, it is important to note that there might be other confounding factors that would create a positive relationship between leverage and forward-looking behavior: For instance, young firms who are usually more leveraged also need to have longer planning horizons and hence might have higher incentives to be more forward-looking. We thank two anonymous referees for suggesting these mechanisms for how leverage and discounting might be related.

Since leverage, size and markups are all correlated, to perform a meaningful test of Predictions 2 and 3 in the data, we use the joint distribution of debt-to-asset ratio and log markups to divide the sample to low leverage and high markup firms vs. the rest of the sample. More specifically, we divide the Compustat sample using the lag of negative debt-to-asset and the lag of log markup. Therefore, for a given cutoff p, we create two samples: (i) one with all observations above the p^{th} percentile of the lag of debt-to-asset $\times -1$ and above the p^{th} percentile of the lag of log markup; (ii) and another with all other observations.[16] This subsampling method guarantees that both Predictions 2 and 3 go in the same direction, and we expect the coefficient on changes in future sales to be larger for the above the p^{th} percentile group. Next, we repeat the IV-GMM estimation in Equation (13) for the resulting two subsamples.

The results of this exercise using 30^{th} percentile as the cutoff is reported in Table 4 and is consistent with Predictions 2 and 3.[17] The coefficient on $\Delta \text{log(Sales)}_{i,t+1}$ for the group above the percentile cutoff is significantly larger than the estimate using all other observations, and more than thrice the size of the coefficient reported in our benchmark regression in Column (3) of Table 3. [TABLES 4 AND 5 HERE.]

We selected 30^{th} percentile as the cutoff to create two subsamples with approximately the same number of observations, but it is important to note that the results are robust to other cutoff selections. Moreover, the results are qualitatively similar if we proxy $1/N$ with a firm's sales relative to their SIC 1-digit industry sales within a year.

[16] Note that if the two variables are positively correlated and we select a cutoff based on one variable only, then Predictions 2 and 3 will go in opposite directions. Thus, the predictions of the models would be ambiguous.

[17] We can do a similar exercise and split the sample based on the lag of relative size rather than the lag of log markup to proxy for $1/N$ in the model, which can stand either for competitiveness or concentration in an oligopoly (See Section 5 for a discussion of this). The results are robust and similar as presented in Table 5.

3.3 Evidence from New Zealand

Our estimation strategy in Compustat relies on estimating markups as well as assuming rational expectations with full information,[18] so that we can instrument for these expectations using the realized values of future sales and markups. In this section, we depart from these assumptions and provide direct evidence for the law of motion for markups using data on a firm's expectations from an expectations survey conducted by Coibion et al. (2018) in New Zealand. In particular, in this survey, firms were asked to provide information about their number of competitors, average markup, current markup, their expected growth in sales, and their next expected price change, which provides an alternative strategy to estimate the law of motion using firms' answers to these questions. Table 6 provides summary statistics of these variables in the data. Using this dataset, we estimate the following specification in the survey data:

$$\hat{\mu}_{i,j} = \phi_1 Ex\Delta Sales_{i,j} + \phi_2 Ex\Delta Price_{i,j} + \lambda_i + \varepsilon_{i,j} \tag{15}$$

where i, j denotes firm j in industry i, $\hat{\mu}_{i,j}$ is the deviation of i, j's markup from its average markup, $ExSales_{i,j}$ is the firm's expected sales growth, $Ex\Delta Price$ is its expected price change and λ_i is an industry fixed effect. Online Appendix F shows how this cross-sectional data allows us to test the predictions of the model. In particular, under fairly regular assumptions signs of ϕ_1 and ϕ_2 correspond directly to signs of ψ_1 and ψ_2 in the model. [TABLE 6 HERE.]

While the predictions of the customer base model hold for any number of competitors (including monopolistic competition), a distinct characteristic of the implicit collusion model is that the coefficients on the law of motion should be larger for firms with a smaller number of competitors. With that in mind, to test the validity of the law of motion, we divide the sample into two sub-

[18] In Online Appendix G we show that while full information rational expectations is a sufficient condition for our law of motion for markups, it is not necessary. In particular, we show that as long as firms' expectations of their *own* future sales growths coincide with full information rational expectations of those sales growths, aggregation works in the sense that firms' [lack of] knowledge about aggregate variables is irrelevant to the derivation of the law of motion for markups.

samples, firms with more than 20 competitors, i.e. competitive firms, and firms with fewer than 20 competitors.[19] Table 7 shows the results of running the regression in Equation (15) for these two subsamples. The results are consistent with the implicit collusion model in several dimensions. [TABLE 7 HERE.]

First, the coefficients on expected sales and on expected price changes are positive and negative respectively for the $N < 20$ sub-sample. Both of these signs are consistent with the implicit collusion model but at odds with the customer-base model, as they suggest that oligopolistic firms (1) increase their markup with higher expected sales, and (2) decrease their markups with positive expected price changes. The positive sign on expected sales growth is consistent with the prediction of the implicit collusion model that oligopolistic firms can sustain higher markups when sales are expected to be higher in the future. In contrast, this positive sign goes against the prediction of the customer base models that firms with higher expected sales in the future should reduce their markups to attract more customers. Moreover, the fact that these coefficients are only significant for firms with fewer competitors, but not for the more competitive firms, provides a placebo test for the implicit collusion model—as it should hold more significantly for less competitive firms.

4 Implications for Cyclicality of Markups

Given that our empirical results support the predictions of the implicit collusion model, in this section we investigate the implications of our law of motion for markups in a calibrated implicit collusion model with supply (TFP) and demand (government spending) shocks by extending the partial equilibrium model from Section 2.1 to a general equilibrium model. Our argument here revolves around the fact that the firms' net present value of future sales growths are closely linked

[19] We also drop firms with less than 2 competitors considering the possibility of a non-binding incentive compatibility constraint for these firms. Figure 4 in the Online Appendix shows how these estimates change as a function of this cutoff. The results are locally robust but the data is very granular, with a substantial number of firms reporting they have a round number of competitors, as shown in Figure 5.

to *output growth* in the economy. Therefore, what determines the cyclicality of markups is the cyclicality of output growth in the general equilibrium.

Households, the Government and Market Clearing. A representative household solves the following standard problem with investment adjustment costs:

$$\max_{\{(C_t,L_t,I_t,K_{t+1},B_t)\}_{t=0}^{\infty}} \mathbb{E}_0 \sum_{t=0}^{\infty} \beta^t \left[\frac{C_t^{1-\theta}}{1-\theta} - \phi\frac{L_t^{1+1/\epsilon}}{1+1/\epsilon}\right]$$

$$s.t. \quad P_tC_t + P_tI_t + B_t \leq W_tL_t + R_tK_t + (1+i_{t-1})B_{t-1} + \int_0^1 \sum_{j=1}^{N} \Pi_{i,j,t}di - T_t$$

$$K_{t+1} = (1-\delta)K_t + (1 - S(\tfrac{I_t}{I_{t-1}}))I_t, \quad S(\tfrac{I_t}{I_{t-1}}) \equiv \tfrac{a}{2}(1 - \tfrac{I_t}{I_{t-1}})^2$$

where C_t is consumption, L_t is labor supply, I_t is investment, K_t is capital, and B_t is a nominal riskless bond with nominal return i_t at $t+1$. The investment adjustment cost is included to allow for a hump-shape in the response of output to shocks. As we discuss later, the extent of this hump-shape response is crucial in determining the cyclicality of markups.

There is also a government that uses lump-sum taxes from households to conduct fiscal policy, G_t. We assume that G_t follows an AR(2) stochastic process

$$G_t = \bar{G}Z_t^g, \quad \log(Z_t^g) = \rho_1^g \log(Z_{t-1}^g) + \rho_2^g \log(Z_{t-2}^g) + \sigma_g\varepsilon_{g,t}, \quad \varepsilon_{g,t} \sim \mathcal{N}(0,1)$$

The AR(2) assumption on the government spending process is to allow for a hump-shaped response of output to a fiscal policy shock. Moreover, we assume that the monetary policy is set based on the Taylor rule $i_t = \phi_\pi \log(P_t/P_{t-1})$. Finally, the market clearing conditions for the final good, capital and labor markets are

$$C_t + I_t + G_t = Y_t, \quad K_t = \int_0^1 \sum_{j=1}^{N} K_{i,j,t}di, \quad L_t = \int_0^1 \sum_{j=1}^{N} L_{i,j,t}di$$

4.1 Calibration and Simulation

In this section, by simulating a log-linearized version of the model around a steady state in which the incentive compatibility constraint binds, and show that the cyclicality of markups is directly linked to how hump-shaped the response of output is to shocks.

Parameters. We set $\beta = 0.993$ to match a steady-state annual real interest rate of 3 percent, $\alpha = 0.35$ to match a steady state share of capital income of 35 percent, $\delta = 0.025$ to match a 10 percent annual rate of depreciation on capital, $\phi = 8$ to match a steady state labor supply of 0.3, $\bar{G} = 0.2$ to match a steady state G/Y of 20 percent, and $a = 2.48$ following Christiano et al. (2005). We also set the response of the central bank to inflation in Taylor rule, ϕ_π to the common value of 1.5. Furthermore, we set the Frisch labor supply elasticity, ϵ, to 2.5. Moreover, we set the elasticity of substitution across sectoral goods, σ, to 4, and the elasticity of substitution within sectoral goods, η, to 20. We set $\gamma = 0.8$ and $N = 15$ to match a steady state markup level of 20 percent.[20]

Although the values of N and γ are calibrated in an arbitrary fashion, we show in a series of robustness checks in Online Appendix C, the model is not very sensitive to reasonable variations in them. The qualitative results in terms of the direction of the cyclicality of markups are robust to any calibration as long as $\eta > \sigma$.

Finally, we set the persistence of the technology shock to 0.95. For the persistence parameters of the government spending shock, we run the following regression on the quarterly data for real government consumption expenditures and gross investment from 1947Q1 to 2014Q1:

$$\log(G_t) = Constant + \rho_1^g \log(G_{t-1}) + \rho_2^g \log(G_{t-2}) + \varepsilon_t$$

which gives the estimates $\rho_1^g = 1.51$ and $\rho_2^g = -0.52$. We also consider alternative persistence parameters for robustness checks in Online Appendix C.

Impulse Response Functions. First, consider the case of no investment adjustment cost ($a = 0$). The dashed curves in Figure 1a show the impulse responses of this model to a 1 percent technology shock.[21] The key observation is that in this setting, output jumps up on impact and converges back to zero as the effect of the transitory shock fades away. Moreover, the response of stochastic discount rate, which is given by $Q_{t,t+1} = \beta \frac{u'(C_{t+1})}{u'(C_t)}$, is countercyclical given that households are able to

[20] Given σ and η, this is the highest level for γ for which the incentive compatibility constraint binds, and the Blanchard Kahn condition for the law of motion for markups holds.

[21] We use Dynare (Adjemian et al., 2011) to solve the model.

smooth their consumption without being restricted by costly investment. The fact that consumption has an inertial response to the technology shock is a crucial element to the countercyclicality of stochastic discount rates. On impact, households expect that their consumption will peak later in the expansion; therefore, they are not really concerned about future states as they know they will have a higher consumption. [FIGURE 1 HERE.]

By Equation 7, the combination of countercyclical output growth and discount rates gives rise to countercyclical markups. The interpretation from the firm side is that on impact, firms know that demand is at its highest. This expectation along with the low price of future profits increases firms' incentives to deviate from implicit collusion and forces the oligopoly to settle on a lower markup in order to eliminate these incentives.

A similar exercise can be done with the government spending shock. Suppose that Z_t^g is an AR(1) process with persistence 0.95. The impulse response functions of the model to such a shock is illustrated by the dashed curves in Figure 1b. On impact, government spending is at its highest, which means that private consumption is at its lowest. First, since private consumption will increase to its steady state level, such a shock would give rise to countercyclical stochastic discount rates. Moreover, the income effect of G is at its highest on impact, so that Y will peak immediately due to a jump in labor supply and converge back to its steady state as the shock fades away. Again, the combination of countercyclical discount rates and output growth yields countercyclical markups.

However, empirical evidence on TFP shocks and government spending shocks suggests that the response of output to these shocks is hump-shaped such that the peak effect happens not on impact but in later periods.[22] To allow for such a response, we introduce investment adjustment costs and an AR(2) process for government spending. Solid curves in Figure 1a show the IRFs of the model to a 1% technology shock when $a = 2.48$. With positive adjustment costs, two things

[22] For empirical evidence on the hump-shaped response of output to productivity and government spending shocks, see, for example, Sims (2011); Smets and Wouters (2007); Ramey (2011); Ramey and Shapiro (1998); Christiano et al. (2005); Monacelli and Perotti (2008); Nekarda and Ramey (2020).

happen. First, investment does not jump on impact and has an inertial response, which translates to a hump-shaped response in output. Second, households now face a stronger trade-off in smoothing their consumption because they face costly investment, which gives rise to procyclical stochastic discount rates. Therefore, on impact firms expect their demand to increase in future periods, which gives them the incentive to sustain higher markups as they expect their demand to increase. Hence, on impact one would expect a higher markup than the steady state, making markups procyclical.

A similar exercise can be done with the government spending shock by assuming an AR(2) process for Z_t^g. Figure 1b depicts the IRFs of the model to such a shock. The hump-shaped implementation of the fiscal policy translates to hump-shaped output and consumption responses, as shown by solid curves in Figure 1b, which in turn produce procyclical markups for similar reasons to the case of the technology shock with $a > 0$.[23]

Matching the Evidence for Conditional Cyclicality of Markups. So far we have shown that a direct implication of our law of motion for markups is that it *reverses* the cyclicality of markups once the model matches the hump-shaped response of output to shocks. Our model predicts that under hump-shaped output response to TFP and government spending shocks, markups are *procyclical*.

These predictions directly relate to and are qualitatively consistent with, the evidence on the conditional cyclicality of markups in recent work by Nekarda and Ramey (2020), who find that (1) output response is hump-shaped with respect to TFP and government spending shocks, and (2) markups are procyclical conditional on these two types of shocks. It is important to note that without the hump-shaped response of output, the model would completely miss these two facts and would deliver the wrong prediction that markups are countercyclical with respect to both TFP and government spending shocks. [FIGURE 2 HERE]

[23] For completeness, we have also included IRFs of a customer-base model in Figure 6 in the Online Appendix, where the calibration is such that the frictionless markup, μ_s, is 11.5%, and the average markup, μ, is 10%. In terms of parameters, the only change compared to the previous section is $\eta = 10$. Notice that in that model markups are countercyclical once the model matches the hump-shaped response of output to shocks.

To formally show this, Figure 2 plots the cross-correlation of markup and output conditional on TFP shocks under the model with and without hump-shaped output response against empirical estimates from Nekarda and Ramey (2020)'s analysis for these correlations.[24] While the model *without* inertia (hump-shaped response of output) fully misses the sign of these correlations, implying that markup and output are negatively correlated, the model *with* inertia gets the sign of and the magnitude of the contemporaneous and lagged correlations of markup and output conditional on TFP shocks right.[25]

It is important to note that while the model matches the contemporaneous and lagged conditional correlations correctly, it predicts that leads of markups should be *countercyclical*, which is not the case in Nekarda and Ramey (2020)'s estimates. This is because in the model markups fall below their steady-state level once output peaks, whereas in Nekarda and Ramey (2020)'s estimates, markups remain procyclical for much longer and their point estimates only fall below their steady state level after 7 quarters. For completeness, Figure 7 in the Online Appendix shows the results of a similar exercise in a customer-base model, and illustrates how this model completely misses the sign and structure of these cross-correlations.

[24] The empirical estimates for these correlations are constructed using replication codes from Nekarda and Ramey (2020)'s TFP SVAR that includes log level of Fernald (2014)'s utilization adjusted measure of TFP, log real GDP per capita, log of the output price deflator, the three months Treasury bill rate and log of Nekarda and Ramey (2020)'s measure of markup based on the labor share of output, allowing for overhead labor. Using the results of this estimation, we calculate the correlations of output and markup conditional on shocks to TFP.

[25] The same exercise can be done with government spending shocks, and while the results qualitatively remain the same – meaning that correlations in the inertial model are larger than in the model with no inertia in the output response – the inertia created by the AR(2) process is not enough to make the conditional correlation positive.

5 Discussion

Strengths and Weaknesses of the Compustat Data. In order to use the production function approach, we need detailed financial data that covers a long period of time for a broad section of the firms in the economy. Compustat fits these criteria very well: It contains yearly data from 1950 to 2016 and it covers 29% of private U.S. employment (Davis et al., 2006). Nonetheless, there is a serious concern that Compustat is not representative of the overall economy as it only contains publicly traded firms. Therefore, a careful discussion of how this issue might affect the external validity of our results is necessary.

Specifically challenging to our paper, Ali et al. (2009) show that measures of industry concentration in the Compustat data are not strongly correlated with measures using the Census data. This concern is relevant to our results because we use measures of concentration or markups to test our model predictions in Section 3, which, in our model, are directly related to both concentration ($1/N$ is the market share of every firm) or competition (firms' markups decrease with N as an oligopoly becomes more competitive). There is a long history regarding the pros and cons of associating market power and concentration, especially that macroeconomic models—including the ones considered here—usually deliver a very tight relationship between the two (see, e.g., Syverson, 2019; Afrouzi et al., 2020). Thus, there is a question of how to map N to the data as both the relative size and competitiveness of an industry might be appropriate proxies to N in the model. However, due to concerns about the weak association of concentration measures between Census and Compustat, we use the lag of log markups as our main proxy for the competitiveness of an industry but show that our results are robust to using measures of concentration too. A finding by De Loecker et al. (2020) that justifies why using lag of markups might not be subject to sample selection issues in Compustat relative to Census is that the estimated markups using Census data for the manufacturing, retail, and wholesale sectors present similar patterns as in Compustat.

There is also a related concern that, while the competition structure in the Compustat data is represented by an oligopoly model due to the high concentration of industries in that dataset, the overall economy might not be, especially since Compustat tends to contain larger and older firms. Thus our results might be contaminated by the more concentrated nature of firms in Compustat, especially if the overall economy is better represented by a model with perfect competition. Nonetheless, evidence from the Census data supports the hypothesis that the overall competition structure of the economy also exhibits oligopolistic traits. For instance, there is a growing literature documenting: (i) a rise in sales concentration within (four-digit) industries and "superstar firms" (Autor et al., 2020; Furman and Orszag, 2018; Ganapati, 2021); (ii) and an increase in profits (Barkai, 2020). While selection to Compustat might affect the quantitative significance of our estimates, the qualitative results hold for an arbitrarily competitive oligopolistic structure: It is only in the limiting case of perfect competition ($N \to \infty$) that these predictions become irrelevant.

Measurement of Markups. In this paper, we use the production-based methodology of De Loecker et al. (2020) for estimating markups because it delivers markups estimates for a broad range of firms over a long enough period of time so that we can take our model predictions to the data—See Online Appendix E for a detailed discussion. However, by now, there is an extensive literature that debates the advantages and disadvantages of using this method.

Markups are notoriously hard to estimate because marginal costs are not observed in the data, and thus any attempt for providing estimates for a broad range of industries needs to systematically address this issue. Production-based methods provide such a systematic approach by attempting to estimate marginal costs from average costs of an arbitrary set of variable inputs (see, e.g., Basu, 2019). By utilizing the optimality conditions for firms' cost minimization, these methods show that marginal costs are proportional to the sales share of an arbitrary set of variable inputs *up to the elasticity of production with respect to these inputs*. Since sales shares of different input costs are observable in the data (barring issues with availability and measurement error, which Basu (2019) discusses in detail), the main challenge for providing consistent estimates of markups is to estimate these production elasticities, which is the main focus of De Loecker et al. (2020). In fact, many of the criticisms to the production-based methodology are related to issues that arise in estimating such production elasticities. For instance, Bond et al. (2021) point out that while the literature usually uses revenue data to estimate these elasticities, such attempts could lead to inconsistent estimates (we discuss these criticisms in more detail in Online Appendix E).

These issues are especially relevant for the estimation of *level of markups* since any systematic bias might lead to systematic over- or underestimation of markups. These criticisms also hold for our use of these methods. However, we believe they are not as severe for our case as we only rely on estimates of *relative* markups: Our estimation of the law of motion only uses the change in log markups on the firm level. In turn, this means that we need to make weaker assumptions to obtain consistency in our estimates. Even if there is a bias in estimating production elasticities since we are interested in the change of the log of markups, we only need to assume that the change in the bias of our elasticities is not correlated with our variables of interest, $\Delta \log\left(\text{sales}_{i,t+1}\right)$ and $\log(\text{M}_{i,t+1})$.

Similarly, most of the literature (including us) impose constant output elasticity within an industry, which could be particularly problematic for estimating the level of markups. For instance, suppose that firm A has twice the revenue share as firm B, but its output elasticity is also half. If the firms are in the same industry, then we would estimate firm A's markup as twice as firm B's, whereas in reality their markups are the same. However, as long as their change in the elasticity is the same, we would correctly estimate the change in log markups.

Relation to New Keynesian Models. By assuming that prices are sticky but marginal costs are not, New Keynesian models create variable markups both across time and across firms. In these models, two forces interact in shaping the cyclicality of aggregate markups: among the fraction of firms that are not resetting their prices, markups are countercyclical because in an expansion their marginal costs rise but their prices stay the same. However, among firms that do reset their prices, they might increase their prices by more than the contemporaneous increase in their marginal costs due to the forward-looking nature of price-setting in these environments. Therefore, the aggregate cyclicality would depend on which force dominates in total. To see this formally, consider the linearized version of the firms side of the textbook New Keynesian model (see, e.g., Galí, 2015):

$$p_t^* = (1-\beta\theta)mc_t + \beta\theta\mathbb{E}_t[p_{t+1}^*], \quad p_t = (1-\theta)p_t^* + \theta p_{t-1} \tag{16}$$

where p_t^* is the log-deviation of reset price for firms that are changing their prices at time t from its steady-state level, mc_t is the log-deviation of the nominal marginal cost of firms (which in our setting is given as a function of aggregate wage and the rental rate of capital in Equation 4) from its steady-state level, p_t is the log-deviation of the aggregate price level from its steady-state level, β is the discount rate and finally, $1-\theta$ is the probability resetting prices at every period. Notice that we can define two notions of markups in this economy: $\mu_t^* \equiv p_t^* - mc_t$ as the markup of the firms that reset their prices at time t and $\mu_t \equiv p_t - mc_t$ as the aggregate markup based on the aggregate price level. Rewriting the equations above in terms of these markups we get:

$$\mu_t^* = \beta\theta\mathbb{E}_t[\Delta mc_{t+1}] + \beta\theta\mathbb{E}_t[\mu_{t+1}^*] \tag{17}$$

$$\mu_t = (1-\theta)\mu_t^* + \theta\mu_{t-1} + \theta\Delta mc_t \tag{18}$$

Notice that Equation (17), which is the *law of motion* for markups among price setters at time t resembles our law of motion for implicit collusion and customer-base models with one major difference: instead of relating markups to changes in the net present value of future sales growths, it relates firms' markups to a discounted average of future changes in marginal cost growths. Therefore, insofar as growth in marginal costs, output growth and stochastic discount rates comove positively in response to a shock, the mechanics of how μ_t^* evolves over the business cycle is akin to the other models considered in this paper. Nonetheless, the dynamics of aggregate markups is more complicated and also depend on the history of prices and marginal costs due to the stickiness of prices among the fraction of firms that are not changing their prices at time t.

Furthermore, the baseline New Keynesian model does not capture the heterogeneity that we observe in the Compustat based on relative size (Predictions 2 in Section 3). In the baseline, New Keynesian model markups of all firms have the same comovement with their future expected marginal cost growths which is uniquely determined by the discount factor and price stickiness ($\beta\theta$). However, it is worth mentioning that the New Keynesian model can also deliver similar predictions through a completely different mechanism as long as price stickiness is positively correlated with measures of concentration, as documented by Carlton (1986).[26] Nonetheless, such heterogeneity must be assumed exogenously in the textbook New Keynesian model as it does not provide a microfoundation for price stickiness.

Surprisingly, models that do microfound price stickiness, such as menu cost models, seem to predict an opposite correlation between price stickiness and concentration: In a menu cost model (e.g. the one in Golosov and Lucas Jr, 2007), firms that are relatively larger have more revenues at stake for not changing their prices and as a result tend to have narrower Ss bands, which in turn leads to higher frequencies of price changes. Since more concentration increases the market share of such firms, the menu cost model associates more concentration with lower price rigidity.

More broadly, while we view our main contribution as highlighting how dynamic strategies in oligopolistic environments affect markup dynamics in microfounded settings, we believe exploring microfoundations of the price stickiness channel, and how they interact with granular competition, is an interesting avenue for future research to study the interactions of these two mechanisms.[27]

On a final note, the standard New Keynesian models take a center stage in the analysis of Nekarda and Ramey (2020), where they find the predictions of these models for the cyclicality of markups conditional on demand shocks to be inconsistent with the empirical evidence. Nekarda and Ramey (2020) argue that the key to matching the evidence for the cyclicality of markups is having more nominal rigidity on wages than prices. By assuming that prices are stickier than wages, the New Keynesian model inevitably leads to markups falling when wages go up due to demand shocks while prices remain below their ideal level due to stickiness.

It is important to note that the same criticism also applies to the types of models that we consider in this paper as we have focused on real price rigidities and have abstracted away from rigidities in the wage market. While the model with a humped-shaped response of output pushes the cyclicality

[26] We thank an anonymous referee for highlighting this mechanism.

[27] See, e.g., Wang and Werning (2020) who in an environment with oligopolistic competition and Calvo pricing deliver sharp analytical results on how price stickiness amplifies real rigidities introduced by oligopoly.

of markups and output conditional on demand shocks in the right direction, this channel, on its own, is not strong enough to reverse the sign of this correlation. A possible remedy for this issue, that would be an interesting avenue for future work, is to extend the models discussed in this paper to settings where firms' market power is also modeled in factor markets (see, e.g., Berger et al., 2022)). Such models would then create rigidities both in prices as well as wages and would create a new mechanism that would allow for a better model fit in response to demand shocks.

Models with Variable Elasticities of Demand. Another set of models that generate variable markups over the business cycle rely on preferences that, in contrast to CES preferences, generate variable demand elasticities along firms' demand curves and populate firms along those demand curves based on heterogeneity in size.

One approach for variable demand elasticities is to use a generalized aggregator (as in Kimball, 1995) that varies demand elasticity according to Marshall's second law of demand—which requires demand elasticity to decrease with relative price along the demand curve. For instance, this approach has recently been used by Edmond et al. (2018) who study the cost of markups in a firm dynamics model. In these models, markups are static but are determined as a function of a firm's demand elasticity given their marginal cost of production. Formally, let $e(p)$ denote the elasticity of demand at relative price p along the demand curve of the firm. Then, the optimal pricing strategy of a firm with real marginal cost mc in these models is to choose p such that $p = \frac{e(p)}{e(p)-1} mc$, which is an implicit equation in the relative price of the firm and determines this relative price as a function of mc. Marshall's second law of demand then implies that firms with higher marginal costs charge higher relative prices *and* lower markups (see, e.g., Lemma 1 in Afrouzi et al., 2020).

Since these models relate markups to *relative* prices, heterogeneity plays a crucial role for their conclusions on how markups change over the business cycle. To see this, notice that in a model with no heterogeneity, by symmetry, all firms have to charge the same price which implies all relative prices are always 1, independent of whether the economy is booming or in a recession. Therefore, in such an economy markups are *always* constant and given by $e(1)/(e(1) - 1)$. It is when there is heterogeneity in size that these models start to shine, and create predictions for the cyclicality of markups as the distribution of prices across the economy starts to change over the business cycle. The implications of these effect for the cyclicality of markups, however, are unexplored to the best of our knowledge and remains a promising area for future work.

Furthermore, variable demand elasticities also arise in nested CES preferences (similar to what we assumed in Equation 1), combined with heterogeneity in productivity (Atkeson and Burstein, 2008; Burstein et al., 2020). Without any dynamic incentives (i.e. implicit collusion or customer acquisition), pricing decisions in these models are static, where firms set a markup over their marginal cost and their markups are determined by their demand elasticity, which in turn depends on the firm's market share within its sector.

Formally, in such a model, the demand elasticity of a firm j in a sector i with market share $s_{i,j}$ can be written as an average of two elasticities of substitution implied by the nested CES system, weighted by the firm's market share as $e(s_{i,j}) = s_{i,j}\eta + (1 - s_{i,j})\sigma$ which then implies the following markup for the firm (Equation 5 in Burstein et al. 2020):

$$\mu(s_{i,j}) = \frac{\eta}{\eta - 1}\left[\frac{1 - (\frac{\eta-\sigma}{\eta})s_{i,j}}{1 - (\frac{\eta-\sigma}{\eta-1})s_{i,j}}\right] \tag{19}$$

Moreover, defining the aggregate markup of a sector as the inverse labor share of that sector implies that the sectoral markup is the harmonic mean of the individual firms' markups, weighted by their market share ($\mu_i^{-1} = \sum_j s_{i,j}\mu(s_{i,j})^{-1}$; Equation 7 in Burstein et al. 2020). Notice now that the sectoral cyclicality of markups depends on two objects: (1) cyclicality of individual firms' markups, and (2) redistribution of sales across the sector during the business cycle. Therefore, the cyclicality of sectoral markups could go against the cyclicality of individual firms' markups within that sector *if there is enough heterogeneity in size*.

To relate our model in Section 2.1 to this framework, we depart from this setting in two ways. First, we consider environments with dynamic incentives which introduce a different law of motion for markups than the static case presented here. These dynamic incentives create intertemporal considerations for firms in choosing their markups. However, for tractability, we do not allow for heterogeneity in size within sectors, as our symmetric equilibria imply that all firms have the same and *constant* market share $(1/N)$ within their sectors. This assumption is of course restrictive but allow for tractability as well as simplified testable predictions that we take to the data. Extending these predictions and models to environments with heterogeneity in market shares would be a natural next step for future work.

6 Conclusion

In this paper, we revisit the implicit collusion and customer-base models and show they both imply a forward-looking law of motion that relates markups to firms' expectations of the net present value of their future sales growths. We then use data on markups and sales from Compustat and survey data on firms' expectations from New Zealand to test this implied law of motion and find the evidence to be in favor of the implicit collusion models.

In a general equilibrium model, we also show that this law of motion reduces the net present value of firms' future sales growth to their expectations of output growth and stochastic discount rates. Because markups are related to the expected output growth, and not to its level, the conditional expectations of firms for the dynamics of output are key in the model for the cyclicality of markups. In particular, if firms expect a hump-shaped response for output during the business cycle, the predictions of these models are reversed.

Lack of sufficiently rich dynamics in output in previous models has led to the conclusion that implicit collusion models lead to counter-cyclical markups. We show that this prediction is overturned once empirically realistic dynamics of output are incorporated into the model, which also helps the implicit collusion model to match the empirical evidence on the dynamic cross correlation of output and markup conditional on TFP shocks.

References

Adjemian, Stéphane, Houtan Bastani, Michel Juillard, Fréderic Karamé, Junior Maih, Ferhat Mihoubi, George Perendia, Johannes Pfeifer, Marco Ratto, and Sébastien Villemot, "Dynare: Reference Manual Version 4," Dynare Working Papers 1, CEPREMAP 2011.

Afrouzi, Hassan, Andres Drenik, and Ryan Kim, "Growing by the Masses: Revisiting the Link between Firm Size and Market Power," *Available at SSRN 3703244*, 2020.

Ali, Ashiq, Sandy Klasa, and Eric Yeung, "The limitations of industry concentration measures constructed with compustat data: Implications for finance research," *Review of Financial Studies*, 2009, *22* (10), 3839–3871.

Atkeson, Andrew and Ariel Burstein, "Pricing-to-market, trade costs, and international relative prices," *American Economic Review*, 2008, *98* (5), 1998–2031.

Autor, David, David Dorn, Lawrence F Katz, Christina Patterson, and John Van Reenen, "The Fall of the Labor Share and the Rise of Superstar Firms," *The Quarterly Journal of Economics*, 2020, pp. 645–709.

Bagwell, Kyle and Robert W. Staiger, "Collusion over the Business Cycle," *The RAND Journal of Economics*, 04 1997, *28* (1), 82–106.

Barkai, Simcha, "Declining Labor and Capital Shares," *Journal of Finance*, 2020, *75* (5), 2421–2463.

Basu, Susanto, "Are price-cost markups rising in the United States? A discussion of the evidence," *Journal of Economic Perspectives*, 2019, *33* (3), 3–22.

Berger, David, Kyle Herkenhoff, and Simon Mongey, "Labor Market Power," *American Economic Review*, 2022, *112* (4), 1147–93.

Bond, Steve, Arshia Hashemi, Greg Kaplan, and Piotr Zoch, "Some unpleasant markup arithmetic: Production function elasticities and their estimation from production data," *Journal of Monetary Economics*, 2021, *121*, 1–14.

Bornstein, Gideon, "Entry and profits in an aging economy: The role of consumer inertia," Technical Report, mimeo 2018.

Burstein, Ariel, Vasco M Carvalho, and Basile Grassi, "Bottom-up markup fluctuations," Technical Report, National Bureau of Economic Research 2020.

Carlton, Dennis W, "The Rigidity of Prices," *The American Economic Review*, 1986, pp. 637–658.

Christiano, Lawrence J, Martin Eichenbaum, and Charles L Evans, "Nominal rigidities and the dynamic effects of a shock to monetary policy," *Journal of political Economy*, 2005, *113* (1), 1–45.

Christiano, Lawrence, Martin Eichenbaum, and Sergio Rebelo, "When Is the Government Spending Multiplier Large?," *Journal of Political Economy*, 2011, *119* (1), pp. 78–121.

Coibion, Olivier, Yuriy Gorodnichenko, and Saten Kumar, "How Do Firms Form Their Expectations? New Survey Evidence," Working Paper 21092, National Bureau of Economic Research April 2015.

_ , _ , and _ , "How do firms form their expectations? new survey evidence," *American Economic Review*, 2018, *108* (9), 2671–2713.

Davis, Steven J., John Haltiwanger, Ron Jarmin, and Javier Miranda, *Volatility and Dispersion in Business Growth Rates: Publicly Traded Versus Privately Held Firms*, Vol. 21 2006.

Edmond, Chris, Virgiliu Midrigan, and Daniel Yi Xu, "How costly are markups?," Technical Report, National Bureau of Economic Research 2018.

Fernald, John, "A quarterly, utilization-adjusted series on total factor productivity," 2014.

Furman, Jason and Peter Orszag, "A Firm-Level Perspective on the Role of Rents in the Rise in Inequality," 2018, pp. 19–47.

Galí, Jordi, *Monetary policy, inflation, and the business cycle: an introduction to the new Keynesian framework and its applications*, Princeton University Press, 2015.

_ **and Mark Gertler**, "Inflation Dynamics: A Structural Econometric Approach," *Journal of Monetary Economics*, 1999, *2.*

Ganapati, Sharat, "Growing Oligopolies, Prices, Output, and Productivity," *American Economic Journal: Microeconomics*, 2021, *13* (3), 309–327.

Gilchrist, Simon, Raphael Schoenle, Jae Sim, and Egon Zakrajšek, "Inflation dynamics during the financial crisis," *American Economic Review*, 2017, *107* (3), 785–823.

Golosov, Mikhail and Robert E Lucas Jr, "Menu costs and Phillips curves," *Journal of Political Economy*, 2007, *115* (2), 171–199.

Gourio, Francois and Leena Rudanko, "Customer capital," *Review of Economic Studies*, 2014, *81* (3), 1102–1136.

Green, Edward J. and Robert H. Porter, "Noncooperative Collusion under Imperfect Price Information," *Econometrica*, 01 1984, *52* (1), 87–100.

Haltiwanger, John and Joseph E Harrington Jr, "The impact of cyclical demand movements on collusive behavior," *The RAND Journal of Economics*, 1991, pp. 89–106.

Kandori, Michihiro, "Correlated demand shocks and price wars during booms," *The Review of Economic Studies*, 1991, *58* (1), 171–180.

Kaplan, Greg and Guido Menzio, "Shopping externalities and self-fulfilling unemployment fluctuations," *Journal of Political Economy*, 2016, *124* (3), 771–825.

Kimball, Miles, "The Quantitative Analytics of the Basic Neomonetarist Model," *Journal of Money, Credit and Banking*, 1995, *27* (4), 1241–77.

Loecker, Jan De, Jan Eeckhout, and Gabriel Unger, "The rise of market power and the macroeconomic implications," *The Quarterly Journal of Economics*, 2020, *135* (2), 561–644.

Monacelli, Tommaso and Roberto Perotti, "Fiscal policy, wealth effects, and markups," Technical Report, National Bureau of Economic Research 2008.

Nekarda, Christopher J and Valerie A Ramey, “The cyclical behavior of the price-cost markup,” *Journal of Money, Credit and Banking*, 2020, *52* (S2), 319–353.

Ottonello, Pablo and Thomas Winberry, “Financial Heterogeneity and the Investment Channel of Monetary Policy,” *Econometrica*, 2020, *88* (6), 2473–2502.

Paciello, Luigi, Andrea Pozzi, and Nicholas Trachter, “Price Dynamics with Customer Markets,” *International Economic Review*, 2018, pp. 413–446.

Phelps, Edmund S and Sidney G Winter, “Optimal price policy under atomistic competition,” *Microeconomic foundations of employment and inflation theory*, 1970, pp. 309–337.

Ramey, Valerie A., “Identifying Government Spending Shocks: It’s all in the Timing*,” *The Quarterly Journal of Economics*, 2011, *126* (1), 1–50.

Ramey, Valerie A and Matthew D Shapiro, “Costly capital reallocation and the effects of government spending,” in “Carnegie-Rochester Conference Series on Public Policy,” Vol. 48 1998, pp. 145–194.

Ravn, Morten, Stephanie Schmitt-Grohe, and Martin Uribe, “Deep Habits,” *The Review of Economic Studies*, 2006, *73* (1), 195–218.

Rotemberg, Julio J. and Garth Saloner, “A Supergame-Theoretic Model of Price Wars during Booms,” *The American Economic Review*, Jun. 1986, *76* (3), 390–407.

_ **and Michael Woodford**, “Markups and the Business Cycle,” in “NBER Macroeconomics Annual 1991, Volume 6” NBER Chapters, National Bureau of Economic Research, Inc, September 1991, pp. 63–140.

Rotemberg, Julio J and Michael Woodford, “Oligopolistic Pricing and the Effects of Aggregate Demand on Economic Activity,” *Journal of Political Economy*, December 1992, *100* (6), 1153–1207.

_ **and** _ , “The cyclical behavior of prices and costs,” *Handbook of macroeconomics*, 1999, *1*, 1051–1135.

Sims, Eric R., “Permanent and Transitory Technology Shocks and the Behavior of Hours: A Challenge for DSGE Models,” 2011. Manuscript.

Smets, Frank and Rafael Wouters, “Shocks and Frictions in US Business Cycles: A Bayesian DSGE Approach,” *The American Economic Review*, 2007, *97* (3), 586–606.

Syverson, Chad, “Macroeconomics and market power: Context, implications, and open questions,” *Journal of Economic Perspectives*, 2019, *33* (3), 23–43.

Traina, James, “Is Aggregate Market Power Increasing? Production Trends Using Financial Statements,” 2018. Manuscript.

Wang, Olivier and Iván Werning, “Dynamic Oligopoly and Price Stickiness,” Technical Report, National Bureau of Economic Research 2020.

Tables and Figures

Table 2: Descriptive statistics for Compustat data

Statistic	Real sales_{it}	% change in $\text{sales}_{i,t+1}$	markup_{it}	% change in $\text{markup}_{i,t+1}$
Mean	1565.40	25.60	1.41	10.51
P 25	27.57	-5.35	1.06	-3.25
P 50	138.06	5.14	1.23	0.03
P 75	648.47	18.76	1.51	3.33
SD	8458.53	841.17	0.79	321.95

Notes: Each column of the table shows the summary statistics of the corresponding variable in the Compustat data. The dataset contains 242,155 observations for 20,252 firms across 67 years (1960-2016). Real sales_{it} for firm i and year t are reported in millions. To calculate markups, we followed the procedure from De Loecker et al. (2020); Traina (2018)–we first estimated time-invariant but industry-specific (SIC 2-digits) output elasticities using the production function estimation method from De Loecker et al. (2020). We then define markup_{it} as output elasticities_j times sales over cost of goods sold (COGS).

Table 3: Law of motion for the U.S.

	(1)	(2)	(3)
	log($\mathrm{M}_{i,t}$)	log($\mathrm{M}_{i,t}$)	log($\mathrm{M}_{i,t}$)
Δ log($\text{sales}_{i,t+1}$)	-0.163***	0.083	0.166**
	(0.010)	(0.085)	(0.083)
log($\mathrm{M}_{i,t+1}$)	0.833***	1.075***	1.078***
	(0.006)	(0.006)	(0.006)
R^2	0.729	0.622	0.551
Year FE	Yes	Yes	Yes
Industry FEs	Yes	No	Yes
Method	OLS	IV-GMM	IV-GMM
N	145,700	145,700	145,700

$*p < 0.10, **p < 0.05, ***p < 0.01$; standard errors are clustered at the firm-level. The first column contains the results of an OLS regression of log(markup_{it}) on Δ log($\text{sales}_{i,t+1}$) and log($\text{markup}_{i,t+1}$). We also include year and industry (SIC 2-digit codes) fixed effects. The second and third columns report the results of the GMM estimator when using four lags of log($\text{sales}_{i,t}$) and log($\text{markup}_{i,t}$) as instruments. The first column report the results of the OLS specification, restricting the observations to the IV-GMM sample used in columns 2 and 3. Our dataset contains 66 industries, 67 years and 242,155 observations.

Table 4: Law of motion for the U.S. split for above and below the 30th percentile of lag log markup and negative of lag of debt-to-asset ratio

	(1)	(2)
	log(M$_{i,t}$)	log(M$_{i,t}$)
Δlog(sales$_{i,t+1}$)	-0.338**	0.600***
	(0.157)	(0.171)
log(M$_{i,t+1}$)	1.092***	1.026***
	(0.009)	(0.013)
R^2	0.577	0.114
Above 30th percentile	No	Yes
Industry FEs	Yes	Yes
Year FEs	Yes	Yes
Markert share	51.08	48.92
N	73085	72615

$*p < 0.10, **p < 0.05, ***p < 0.01$; standard errors are clustered at the firm-level. We report the GMM estimator of the effects of Δ log(sales$_{i,t+1}$) and log(markup$_{i,t+1}$) on log(markup$_{i,t}$) using four lags of log(sales$_{i,t}$) and log(markup$_{i,t}$) as instruments. We also include year and industry (SIC 2-digit codes) fixed effects. The second column reports the results for observations above the 30th percentile of lag log markup and negative of lag debt-to-asset ratio, while the first column reports the results for all other observations.

Table 5: Law of motion for the U.S. split for above and below the 30^{th} percentile of lag relative sales and negative of lag debt-to-asset ratio

	(1)	(2)
	log($\text{M}_{i,t}$)	log($\text{M}_{i,t}$)
Δlog($\text{sales}_{i,t+1}$)	0.105	0.220^{**}
	(0.071)	(0.088)
log($\text{M}_{i,t+1}$)	1.104^{***}	1.037^{***}
	(0.008)	(0.006)
R^2	0.484	0.699
Above 30^{th} percentile	No	Yes
Industry FEs	Yes	Yes
Year FEs	Yes	Yes
Market share	27.88	72.11
N	71999	73701

$*p < 0.10, **p < 0.05, ***p < 0.01$; standard errors are clustered at the firm-level. We report the GMM estimator of the effects of Δ log($\text{sales}_{i,t+1}$) and log($\text{markup}_{i,t+1}$) on log($\text{markup}_{i,t}$) using four lags of log($\text{sales}_{i,t}$) and log($\text{markup}_{i,t}$) as instruments. We also include year and industry (SIC 2-digit codes) fixed effects. The second column reports the results for observations above the 30^{th} percentile of lag relative sales and negative of lag debt-to-asset ratio, while the first column reports the results for all other observations. We define relative sales as the total sales for a firm in a given year, divided by the total sales of the industry (SIC 1-digit) in that same year.

Table 6: Descriptive statistics for New Zealand data

Statistic	Firm production value	Expected % change in sales	Markup	Expected size of next price change
Mean	992,136.69	4.39	25.02	4.64
P 25	211,600.00	0.00	15.00	2.00
P 50	369,500.00	5.00	25.00	5.00
P 75	1,061,500.00	8.50	34.50	8.00
SD	1,935,235.80	5.84	12.18	5.27

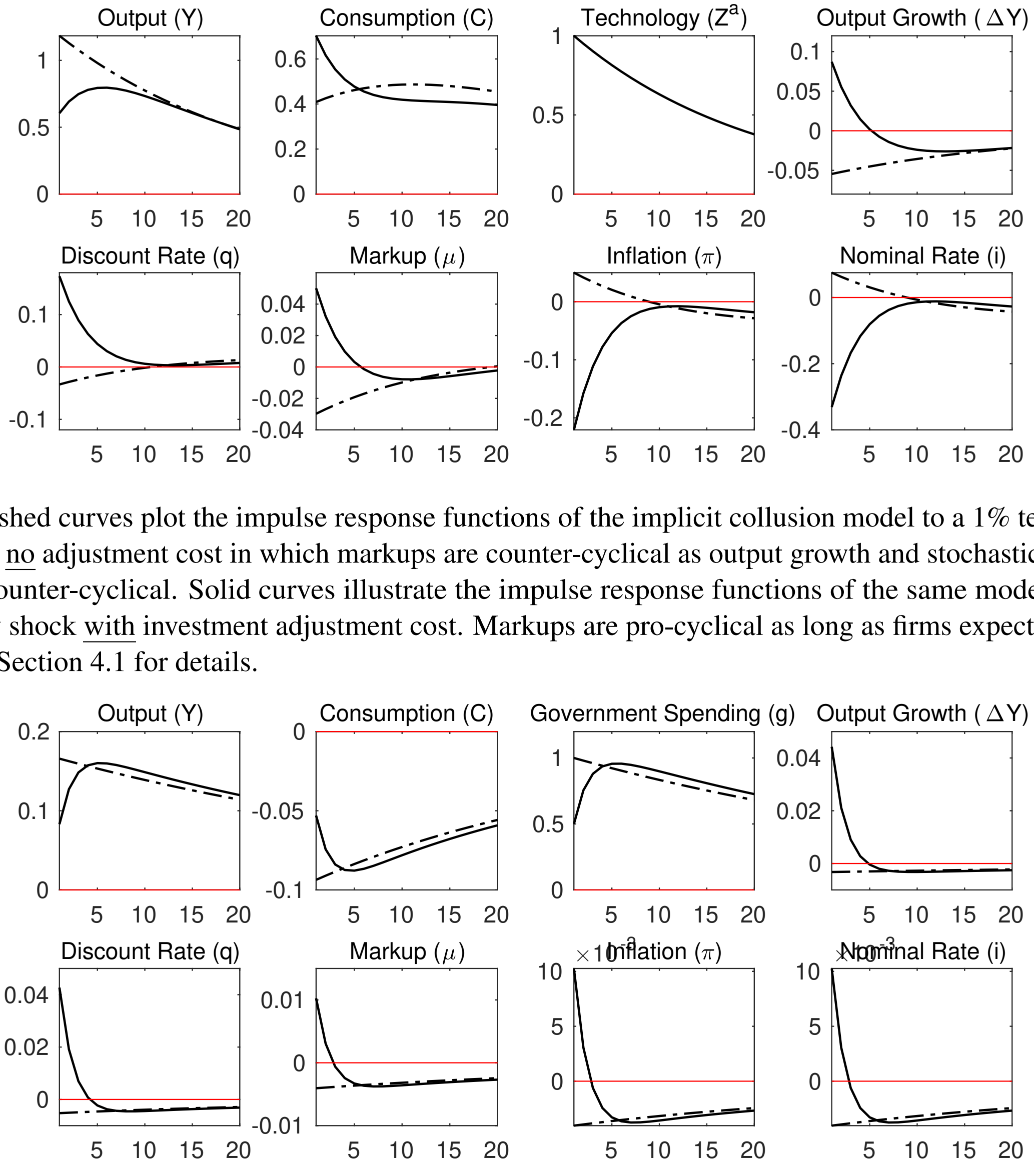
The table provides summary statistics for the survey of firms' expectations from New Zealand. The dataset contains 3,153 observations for 3,153 firms in 18 industries and corresponds to the first wave of the survey conducted by Coibion et al. (2015).

Table 7: Law of Motion for Markups: Survey Data from New Zealand

	(1) Markup	(2) Markup
Expected size of next price change	-0.178***	0.025
	(0.062)	(0.090)
Expected growth in sales	0.163**	-0.037
	(0.082)	(0.107)
R^2	0.118	0.153
Industry FE	Yes	Yes
Number of competitors	$2 \leq$ competitors ≤ 20	$2 > 20$
N	495	200

$*p < 0.10, **p < 0.05, ***p < 0.01$; robust standard errors. The table reports the coefficients for the regression specified in Equation (15) allowing for industry fixed effects. The first column reports the coefficients for firms that report less than 20 but more than 2 competitors. Second column reports the coefficients for firms that report more than 20 competitors.

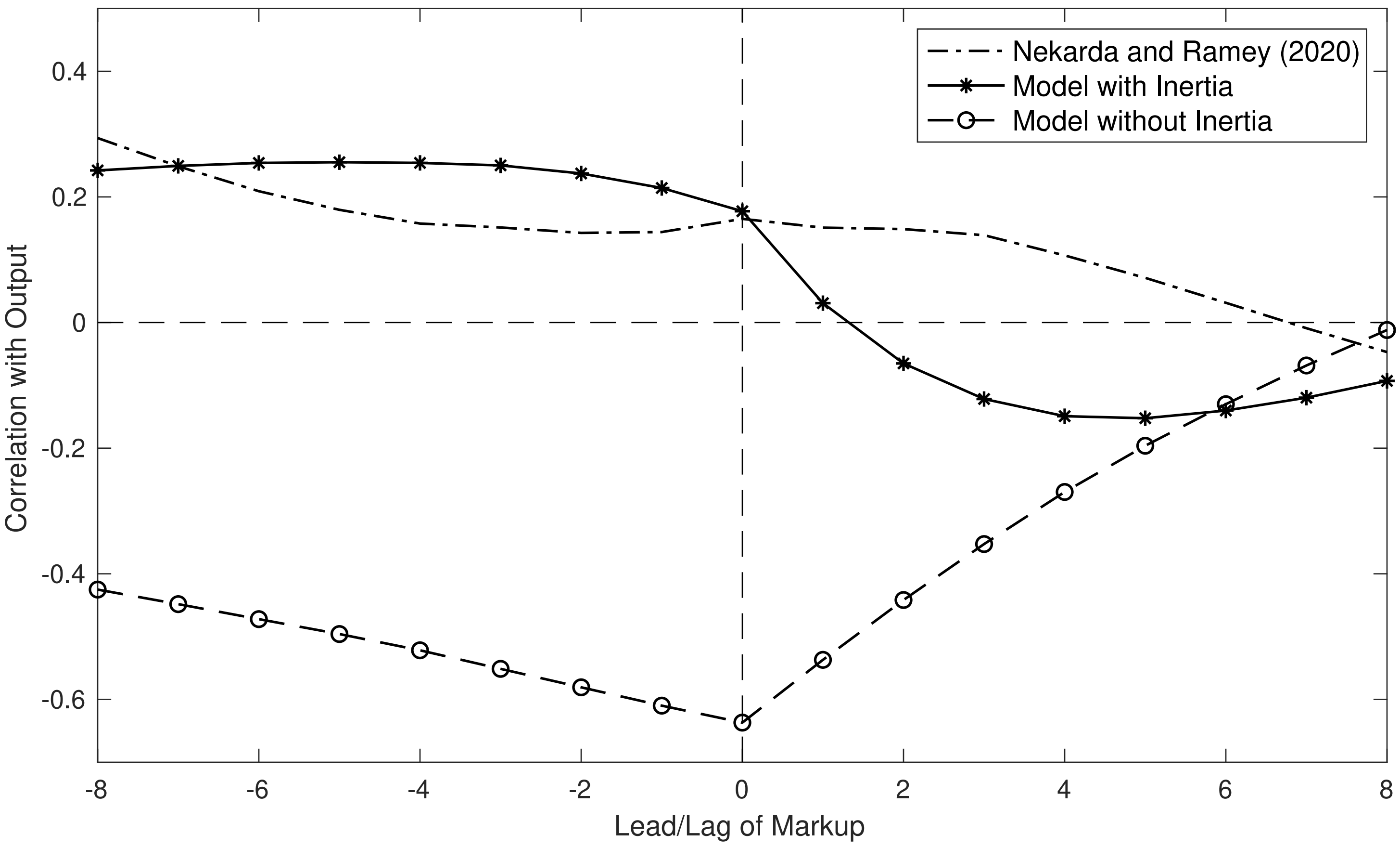
Figure 1: Impulse Response Functions: Implicit Collusion Model



(a) The dashed curves plot the impulse response functions of the implicit collusion model to a 1% technology shock with no adjustment cost in which markups are counter-cyclical as output growth and stochastic discount rates are counter-cyclical. Solid curves illustrate the impulse response functions of the same model to a 1% technology shock with investment adjustment cost. Markups are pro-cyclical as long as firms expect output to grow. See Section 4.1 for details.

(b) The dashed curves plot the impulse response functions of the implicit collusion model to a 1% government spending shock without inertia in which markups are counter-cyclical as output growth is negative during the expansion. Solid curves illustrate the impulse response functions of the same model to an inertial government spending shock that peaks at 1%. Markups are pro-cyclical on impact as output growth and stochastic discount rates are pro-cyclical. See Section 4.1 for details.

Figure 2: Cross-correlation of Markup and Output in the Implicit Collusion Model



Notes: The black curve with square markers depicts correlation of μ_{t+j} with Y_t from the simulated implicit collusion model without inertial response of output conditional on TFP shocks. The dotted curve shows cross-correlation of the cyclical components of markups with real GDP from Nekarda and Ramey (2020)'s analysis for TFP shocks. The black curve with circle markers illustrate this cross-correlation from the simulated implicit collusion model with inertial response of output conditional on TFP shocks. Inertia is crucial in matching the positive correlation between output and markups. See Section 4.1 for details.