# Monopsony Power and the Transmission of Monetary Policy*

Bence Bardóczy† Gideon Bornstein‡ Sergio Salgado§

November 12, 2025

**Abstract**

This paper studies how labor market power affects the transmission of monetary policy. Using administrative U.S. Census data, we show that firms with high monopsony power—those accounting for over 10 percent of the local wage bill—respond less to monetary policy in terms of their wage bill and employment. We then develop a New Keynesian model with heterogeneous firms and oligopsonistic competition to interpret these findings. Wage stickiness combined with firms' labor market power is key to generating the heterogeneous responses that we document. Our model highlights two channels through which oligopsony shapes the aggregate effects of monetary policy: partial passthrough and misallocation. Calibrated to U.S. labor markets, the model implies that the decline in labor market power since the 1980s has increased the output response to monetary policy by about 50 percent.

*We thank Chiara Maggi for her early contributions and extensive discussions that shaped the direction of this paper. For helpful comments, we thank Adrien Auclert, Hassan Afrouzi, Salomé Baslandze (discussant), Luigi Bocola, Émilien Gouin-Bonenfant, Hess Chung, Matias Covarrubias, Lu Han, Sebastian Heise, Kyle Herkenhoff, Oleksiy Kryvtsov, Claudia Macaluso, Simon Mongey, Maarten De Ridder, Erick Sager, Sebastian Di Tella, Christopher Tonetti, Christian Wolf, as well as participants at various seminars and conferences. Tanvi Jindal and Paige Stevenson provided superb research assistance. Any opinions and conclusions expressed herein are those of the authors and do not necessarily represent the views of the U.S. Census Bureau, the Federal Reserve Board of Governors, or the Federal Reserve System. All results have been reviewed to ensure that no confidential information is disclosed.

†Federal Reserve Board of Governors; bence.a.bardoczy@frb.gov

‡The Wharton School, University of Pennsylvania; NBER; gideonbo@wharton.upenn.edu

§The Wharton School, University of Pennsylvania; ssalgado@wharton.upenn.edu

# 1 Introduction

Over the past four decades, local labor markets in the U.S. have become less concentrated.[1] The Herfindahl-Hirschman index of local labor market wage bill has declined from 0.35 in 1980 to 0.24 in 2019.[2] Despite this decline, local labor markets are still highly concentrated: less than 6 percent of the firms in a local labor market account for 40 percent of employment. This high degree of concentration can lead to a partial and heterogeneous pass-through of demand shocks to wages, a key element in the transmission of monetary policy to real outcomes.

In this paper, we study how labor market power shapes the transmission of monetary policy. We address this question both empirically and theoretically. On the empirical side, we use administrative U.S. Census data to document the heterogeneous response of firms to a monetary policy shock. We classify firms into low- and high-monopsony-power groups, according to their wage-bill share in the local labor market. Our analysis shows that low-monopsony-power firms are significantly more responsive to monetary policy shocks both in terms of their wage bill and their employment.

On the theoretical side, we construct a heterogeneous agent New Keynesian model with oligopsonistic competition. The economy consists of a continuum of local labor markets, within each a finite number of heterogeneous firms employ workers. These firms compete à la Bertrand and can adjust their nominal wages infrequently. Our model implies two channels through which oligopsonistic competition affects the aggregate effects of monetary policy: a *partial passthrough channel* and a *misallocation channel*. We develop a numerical method to solve the model, and calibrate it to match key characteristics of the U.S. labor market. In the calibrated model, like in the data, high-monopsony-power firms are less responsive to monetary policy. Quantitatively, we find that oligopsonistic competition damps the effect of monetary policy on output by 24 percent, relative to an economy where firms maintain stable markdowns. Furthermore, the decline in labor market concentration since 1980 has amplified the transmission of monetary policy to aggregate output by more than 50 percent.

The dataset we use to study the heterogeneous response of firms to monetary policy is the Census' Longitudinal Employer-Household Dynamics (LEHD) merged to the Longitudinal Business Database (LBD). The LEHD contains quarterly information on wage bill and employment of all establishments with employees in the U.S. between 1993 to 2022, including the county in which they are located, their age, industry, and the firm to which the establishment belongs. We follow Yeh et al. (2022) in defining a local labor market to be all workers working for firms in the same subsector (3 digit NAICS) in the same county. We classify a firm-county-subsector triplet as a high-labor-market-power firm if it accounts for more than 10 percent of the total wage bill in the local labor market. Otherwise, it is classified as a low-labor-market-power firm.

We use the local projection method (Jordà, 2005) to identify the dynamic effect of monetary policy on firms' wage bill. Our main empirical finding is that low-labor-market-power firms respond more to a change in monetary policy. In response to a 25 basis points reduction in the federal funds rate, the wage bill of low-monopsony-power firms responds by 0.5 log point more than that of high-monopsony-power

[1] See Rossi-Hansberg et al. (2021) and Berger et al. (2022).

[2] Authors' calculation using the U.S. Census Longitudal Business Database. We follow Yeh et al. (2022) in defining a local labor market as a subsector (3-digit NAICS) in a county.

firms in the second year following the shock. To put this number in perspective, while both group of firms expand wage bill following an expansionary monetary policy shock, low-labor-market-power firms response is about 50 percent larger. This heterogeneous response holds even after controlling for the national size of firms as well as their age, changing the definition of a local labor market, and considering different types of monetary policy shock series. Finally, we decompose the wage bill response into changes in employment and changes in the wage, and find that employment accounts for roughly one third of the overall wage bill change.

The heterogeneous response we find across firms suggests that presence and observed changes in labor market power over time can affect the transmission of monetary policy to real outcomes. To gauge at the aggregate effects of labor market power on monetary policy efficacy, we construct a structural general equilibrium model with heterogeneous oligopsonistic firms and nominal rigidities. The labor market structure builds on Berger et al. (2022). The economy consists of a continuum of local labor markets. Each local labor market contains a finite number of firms who differ along their productivity level. Workers in the economy have idiosyncratic tastes towards different local labor markets and towards different firms, resulting in a finite labor supply elasticity for each firm which endogenously varies with the wage bill share of the firm in the local market.

The oligopsonistic firms all produce a homogeneous good but set their wages taking into account the labor supply function they face. They compete à la Bertrand and can adjust their wage infrequently. When a firm adjusts its wage, it internalizes its effect on the local wage index, taking as given the wage strategies of its competitors. We study the equilibrium with commitment, which allows us to accommodate rich heterogeneity across firms and local labor markets. The optimal markdown equals to the average inverse labor supply elasticity until the next wage adjustment, with each period weighted by the probability that the firm remains unable to change its wage. Since labor supply elasticity is lower for firms with a larger labor market share, these large firms set a larger markdown relative to competitive wages. The rest of our model is standard.

Before quantifying the model, we use a special case of the model to show the importance of wage stickiness in driving the heterogeneous response across firms. We consider the limit case where the discount factor is equal to zero, so that optimal wage setting is static. We study the effect of a rise in demand, induced by an expansionary monetary policy, on a local labor market consisting of three firms: a low-, medium-, and high-productivity firms. We first show that absent wage stickiness, all firms adjust their wage by the same magnitude, resulting in a homogeneous response of labor and wage bill despite the heterogeneity in labor market power. When all firms can adjust their wage, their labor market share remains unchanged, they do not change their markdowns, so there is full passthrough from the demand shock to the wage. We then consider the case in which one firm cannot adjust its wage. When that is the case, there is a heterogeneous response between the remaining two firms. With one firm stuck at their previous nominal wage, the labor share of the other two firms goes up. Because the low-productivity firm faces a higher superelasticity of labor supply, its passthrough from prices to wages is higher, and it responds more both in terms of employment and wage bill. Thus, wage stickiness is a key driver of the heterogeneous response across firms.

The stylized model uncovers the two channels through which imperfect competition in the labor market damps the effect of monetary policy on output. First, oligopsonistic competition reduces the passthrough from prices to wages. This is because firms optimally increase their markdowns when demand goes up. The lower passthrough to wages reduces the overall employment response, and as a result, the output response. This is the *partial passthrough channel*. Second, the heterogeneous response across firms affects aggregate total factor productivity (TFP). An expansionary monetary policy shocks increases the employment share of low-monopsony-power firms. These firms are less productive on average, so such shift reduces aggregate TFP. As a result, oligopsonistic competition damps the effect of monetary policy on output by moving TFP in the opposite direction. This is the *misallocation channel*.

Solving the full version of the model brings two new challenges relative to a standard heterogeneous agent New Keynesian model such as Kaplan et al. (2018). First, the finite number of firms operating in each local labor market implies that a law of large numbers cannot be invoked to eliminate the idiosyncratic risk of wage stickiness. Second, firms and workers interact in a large number of markets and the equilibrium wages are determined implicitly from firms' mutual best responses. To overcome these challenges, we adapt the sequence-space Jacobian approach (Boppart et al. 2018; Auclert et al. 2021) in two ways. First, we extend the perfect foresight assumption from including only aggregate variables to also including the idiosyncratic wage adjustment shocks. Second, we introduce an inner layer where we solve for the equilibrium in each local labor market in isolation. This numerical approach allows us to handle many heterogeneous firms across many local labor markets. The estimated model consists of approximately one million firms across 25 thousand labor markets.

We calibrate the model to match key features of the U.S. labor market. Particularly, we target the employment and population share of high-monopsony-power firms, the average HHI across local labor markets, as well as the average markdown. The model is able to replicate the heterogeneous response across low- and high-labor-market-power firms. In response to an expansionary monetary policy shock, the wage bill of low-monopsony-power-firms responds by roughly 60 percent more than that of high-monopsony-power firms.

Labor market power affects the transmission of monetary policy through the partial and heterogeneous passthrough of prices to wages by the oligopsonistic firms. To quantify the impact of oligopsonistic competition on the transmission of monetary policy, we consider a counterfactual economy in which firms have full passthrough.We find that the response of output to a monetary policy shock in the benchmark economy is 24 percent lower than in the full-passthrough counterfactual economy. Two thirds of the difference is attributable to the partial passthrough channel and one third to misallocation.

Finally, we recalibrate the distribution of firms to represent the higher level local concentration that was present in U.S. labor markets in 1980. In this economy, firms face fewer local competitors and have higher market shares, on average. Our analysis suggests that a monetary policy shock is more than 50 percent more effective at stimulating aggregate output in 2019 than it was back in the 1980. A related implication is that the decline in local concentration has flattened the slope of the Phillips curve over time.

**Literature review.** This paper contributes to empirical and theoretical research on labor market power, firm heterogeneity, and monetary policy.

A recent literature leverages administrative data to estimate the degree of labor market power and the evolution of local labor market concentration over time (Azar et al., 2020; Benmelech et al., 2022; Rinz, 2022; Chan et al., 2023). Our definition of a local labor market follows Yeh et al. (2022). We contribute to this empirical literature by establishing a connection between local labor market concentration and the transmission of monetary policy. By showing that firms respond differentially to monetary policy shocks according to their local labor market share, our paper relates to the burgeoning literature that studies the role of firm heterogeneity in monetary policy. This literature has emphasized the role of heterogeneity in size and financial constraints (Gertler and Gilchrist, 1994; Caglio et al., 2021; Champagne and Gouin-Bonenfant, 2022), age (Durante et al., 2022; Cloyne et al., 2023), and default risk (Ottonello and Winberry, 2020). To this list, we add labor market power as a dimension of heterogeneity that modulates firms' responses to monetary policy.

Modeling firm-side labor market power in general equilibrium is typically done in one of two ways. Search and matching frictions, as in Burdett and Mortensen (1998) and Manning (2003), or preference heterogeneity as in Card et al. (2018) and Lamadon et al. (2022). Both of these reasons for labor market power have been explicitly recognized by Robinson (1933) who introduced the term "monopsony". We adopt a preference heterogeneity-based approach, extending the setup of Berger et al. (2022) with a nested CES labor supply system to nominal wage stickiness. In the resulting model, heterogeneous firms play a dynamic game with their local competitors on a large number of markets. Solving this model numerically is challenging. We develop a novel application of the sequence-space Jacobian method of Auclert et al. (2021) that scales efficiently to heterogeneous firms and markets.

Finally, our paper is related to the literature on monetary policy with oligopolistic product markets. Mongey (2021) and Ueda (2023) study duopolies, while Wang and Werning (2022) and Alexander et al. (2024) study oligopolies with homogeneous firms. Compared to these papers, we focus on the labor market and allow for firm heterogeneity in productivity and number of local competitors. Partial passthrough from costs to prices (or in our case, prices to wages) is a central element in all of these models. However, our framework with heterogeneous productivity also features an additional misallocation channel which accounts for about one third of the total impact of oligopsony. Conceptually, this "supply-side channel" of monetary policy is similar to Baqaee et al. (2024) and Meier and Reinelt (2024), who get endogenous fluctuations in aggregate TFP due to reallocation between firms with different desired markups but uniform productivity.

**Layout.** The paper proceeds as follows. In section 2, we establish empirically that firms respond differentially to monetary policy shocks depending of the degree of their labor market power. In section 3, we present our heterogeneous agent New Keynesian model with oligopsonistic labor markets. In section 4, we provide intuition using a special case of our model. In section 5, we introduce a numerical method to solve the model and quantify the effect of oligopsonistic on the transmission of monetary policy. Section 6 concludes.

## 2 Empirical Analysis

In this section, we study whether monetary policy shocks have a differential impact on firms depending on their degree of labor market power. To this end, we combine two administrative data sources from the US Census Bureau. We use these data to construct firm-level measures of employment and wage bill within narrow labor markets and by firm characteristics such as age, firm size, and industry. We complement these data with measures of monetary policy shocks and other macroeconomic time series.

### 2.1 Data

Our primary source of data is the Census' Longitudinal Employer-Household Dynamics (LEHD) which is an employer-employee matched panel dataset constructed from unemployment insurance records. The LEHD contains quarterly data on employment and wages at the individual level and includes information on the worker (e.g., age, gender, and education level) and their employer (e.g., firm identifier, location of the job, and narrow industry groups).

In our analysis, we use information from 1993Q1 to 2021Q4 for all states available at a particular quarter. This implies that the number of states increases over time—from 6 to 23—as new states are included in the sample.[3] The unit of observation in LEHD is an individual unemployment insurance account, which can be associated with multiple firms if the worker is employed in different firms in a given quarter.

We aggregate the wage bill and employment data in LEHD at the firm-county level within a quarter, summing all the wage bill and employment across all the establishments a firm has in a particular county. Importantly, although for the wage bill we record all the wages paid by a firm in a period, part of this wage bill is due to short employment spells that could last even a day. Hence, to reduce the volatility of firm-level employment, we consider a worker to be employed in a particular firm if the individual earned more than 950 dollars within the quarter which corresponds to working full time for at least 3 weeks on the federal minimum wage. Finally, we drop firm-quarter observations with zero wage bill. This leaves us with about 207 million firm-labor market-quarter observations which we use in our main empirical analysis.

We define a local labor market to be all firms within a subsector (3-digit NAICS) in the county. Then, for each firm in a local labor market, we calculate the share of the labor market wage bill it accounts for. We classify firms into low and high labor market power categories. A firm is considered to have high labor market power if it accounts for 10 percent or more of the total wage bill within the local labor market. All other firms within a labor market are classified as having low labor market power. Notice that because we aggregate wage bill and employment of firms across subsectors and counties, a firm present in multiple labor markers can be classified as having high labor market power in one labor market (e.g., a grocery chain is the only supermarket in a small town) but as having low labor market power in a different labor market (e.g., the same chain but in a big city).

[3] More precisely, we have information on 6 states in 1993, including California. By 2000, this increased to 21, and to 23 states in 2004. By the end of our sample, we have the following states: Arkansas, Arizona, California, Connecticut, Colorado, Kansas, Connecticut, Maine, South Carolina, North Dakota, New Jersey, New Mexico, Nevada, Ohio, Oklahoma, Pennsylvania, South Dakota, Tennessee, Utah, Virginia, Washington, Wisconsin, Wyoming.

We combine the data from LEHD with information from the Longitudinal Business Database (LBD), which contains annual information on wage bill and employment for all firms in the private sector from 1978 to 2022 and revenue information starting in 1998. The key advantage of LBD relative to LEHD is its coverage, which spans the entire private business sector for all states. We use LBD to obtain the age of the firm, the levels of employment, wage bill, and revenue at the national and at the local labor market levels. Furthermore, the geographical information contained in LBD allows us to define a local labor market using Commuting Zones rather than counties. We use this information in our robustness exercises.

We combine this dataset with quarterly macroeconomic aggregates (e.g., GDP and unemployment) and a measure of monetary policy shocks obtained from Jarociński and Karádi (2020).

## 2.2 Econometric Specification and Results

To measure the impact of monetary shocks on the labor market and how that varies depending on the degree of firms' labor market power, we run a series of local projections (Jordà, 2005) of the form

$$\tilde{y}_{i,t+h} - y_{i,t-1} = \delta_{t,s,g,h} + \beta_h \varepsilon_t^{MP} \times I\left[s_{i,t-1} \leq 10\%\right] + \Gamma_h \mathbf{X}_{i,t-1} + \epsilon_{i,t+h}, \quad \text{for } h = \{0, \ldots, 15\} \tag{2.1}$$

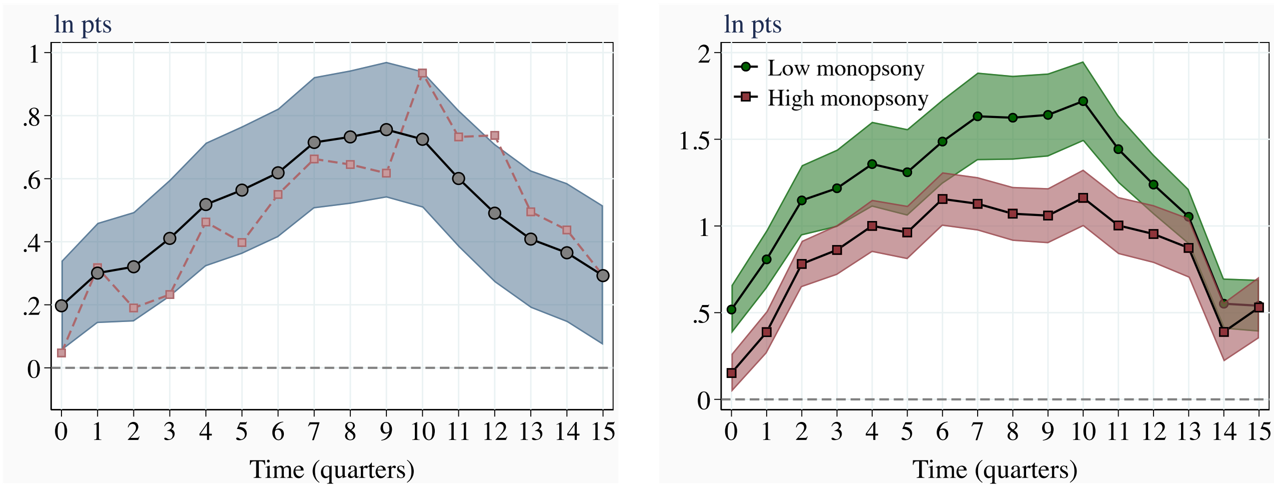
where $\tilde{y}_{i,t+h} = \frac{1}{4}\sum_{\tau=0}^{3} y_{i,t+h+\tau}$ is a forward-looking moving average of the log of the wage bill for firm $i$ in quarter $t$, and $y_{i,t-1}$ is the log of the wage bill in period $t-1$. This simple moving-average local projection specification, which is motivated by the smooth local projection approach of Barnichon and Brownlees (2019), has the advantage that it can be easily implemented with large panel datasets and smooths out the large jumps of wage bill and employment that are usually observed in firm-level data. We also include time-subsector-monopsony power group fixed effects $\delta_{t,s,g,h}$ to control for other aggregate effects that could impact firm's wage bill within an industry. The matrix $X_{i,t}$ contains aggregate controls (CPI inflation, unemployment rate, and GDP), firm-level controls (four lags of $y_{i,t}$), and state-by-subsector fixed effects. Finally, the coefficient $\beta_h$ capture the response of the wage bill of a firm with low monopsony power (identified by the dummy $I\left[s_{i,t-1} \leq 10\%\right]$) to a monetary policy surprise, $\varepsilon_t$, *relative* to high-monoposony-power firms.

Our baseline results are displayed in Figure 1, which shows the response of the log wage bill to an unexpected decrease in the federal funds rate of 25 basis points. The dotted line in panel 1a shows a positive and statistically significant response of the wage bill to an expansionary monetary shock. For comparison, we show the results from a standard panel local projection (dashed line). We find that on average, the wage bill increases by 0.5 log point more for low-monopsony-power firms relative to high-monopsony-power firms four quarters after the shock. The relative effect reaches about 0.9 log point ten quarters after the shock.

It is possible that the larger response of low-monopsony-power firms depicted in Figure 1a only implies more churning of workers from high to low-monopsony-power firms without any change in aggregate employment. To verify this, Figure 1b displays the within-group response of the wage bill for high- and low-monopsony-power firms. Two results are worth noticing. First, both groups respond positively to an expansionary monetary shock; however, firms with low monopsony power respond

Figure 1: Wage Bill Response to A Monetary Policy Shock

(a) Relative Response (b) Response Across Groups




The figure shows the relative response of firm-level log wage bill (Panel a) and within the group of high and low monopsony power (Panel b). We plot the coefficient $\beta_h$ for each horizon $h$ from 0 to 15. To reduce the impact of outliers, we trim the sample for each horizon $h$ at the bottom and top 0.1 percent of the distribution of the change wage bill or employment between periods $t$ and $t+h$. The shaded areas represent 95 percent confidence intervals. Standard errors clustered at the local labor market level. Regressions include an average of 155 million firm-quarter observations

relatively more. Second, the response of high-monopsony-power firms is about 0.9 log point after four quarters, whereas that of low-monopsony-power firms is about 1.4 log points, indicating that low-monopsony-power firms respond 50 percent more than high-monopsony-power firms.

## 2.3 Robustness

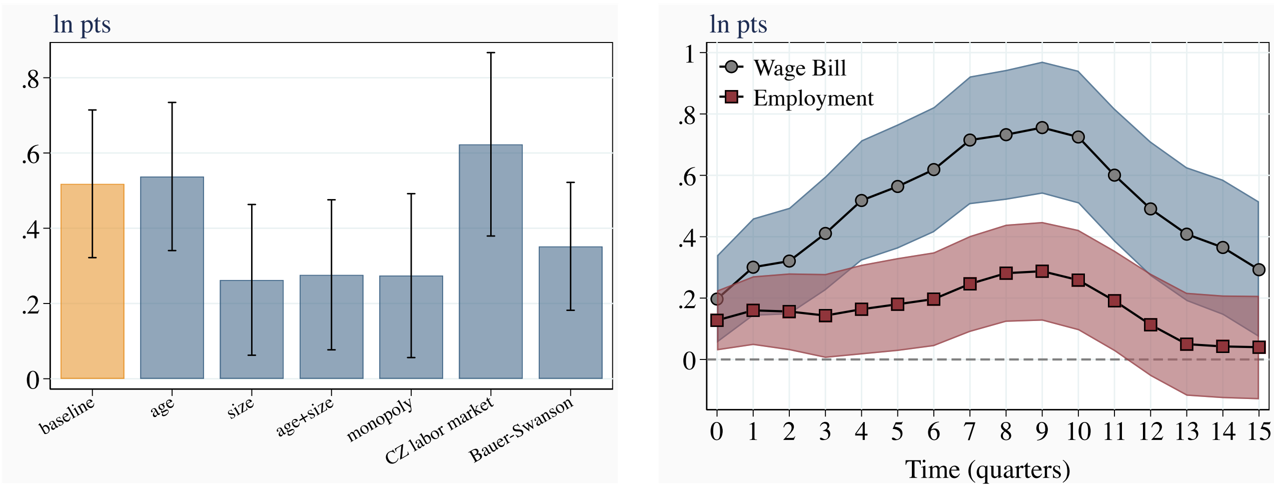
Low-monopsony-power firms tend to be smaller and younger than high-monopsony-power firms, which could make them more responsive to monetary policy shocks. To address this, we consider a list of robustness exercises which are summerized by Figure 2a. This figure shows how our headline result—the response of the wage bill for low-monopsony-power firms relative to high-monopsony-power after a 25 basis points expansionary monetary surprise four quarters after the shock—as we change the specification.[4] The lines represent 95 percent confidence intervals. In each case, we interact the monetary policy shock with a dummy capturing a particular firm characteristic.

**Age.** We divide our firms by age at the moment of the shock into three groups—less than 5 years, between 5 and 10, and above 10 years old—and we plot the relative response of low versus high-monopsony-power firms. The second bar in Figure 2a shows that controlling for age does not change our results. This is because we do not observe a significantly larger response for young (less than 5 years old) or middle-aged firms (between 6 and 10) relative to old firms (10+ years old) up to eight quarters after the shock (Appendix Figure C.1).

[4] The full moving average impulse response is shown in Appendix Figure C.3.

Figure 2: Wage Bill Response to A Monetary Policy Shock

(a) Robustness (b) Employment




**Notes**: Figure 2a shows the wage bill response of low-monopsony-power firms relative to high-monopsony-power firms to a 25 basis points expansionary monetary policy shock at 4 quarters after the shock, for different specifications. In each case, we include the interaction of the corresponding control and the monetary policy shock. Bars represent 95 percent confidence interval. *Age* indicates firm age controls. *Monopoly* is a dummy equal to 1 if the firm is at the top 5 percent of the within NAICS-4 sales distribution. *Size* indicates firm employment size controls. *CZ labor market* indicates labor market power defined at the commuting zone level. *Bauer-Swanson* indicates Bauer and Swanson (2023) monetary policy shocks. Figure 2b shows the response of employment in the baseline specification.

**Size.** We proceed similarly when controlling for employment size. That is, we divide firms within a labor market by (lagged) employment in four groups—less than 5 workers, between 6 and 10, between 10 and 20, and above 20 workers. The results, shown in the second bar of Figure 2a, indicate that, conditional on firm size, low-monopsony-power firms respond by about .25 log point more than high-monopsony-power firms. This is because the response of small and medium-sized firms within a labor market tends to be stronger than that of large firms (right panel of appendix Figure C.1), which tend to have higher monopsony power, thus reducing the importance of firm relative size within a local labor market. We find similar results when controlling simultaneously for firm age and size or if we define firm size at the national rather than at the local labor market level (Appendix Figure C.2).

**Monopoly power.** We then study whether controlling for the relative share of sales within a subsector—which can be interpreted as a measure of firms' monopoly power—changes our results. In this case, we rely on information from LBD since 1998 and define a firm to have high monopoly power if the firm has more than 5 percent of the revenue share at the national level within a NAICS-4 subsector. The results are shown in Figure 2a, in the column labeled "monopoly". Similarly to the case when we control by size, adding monopoly controls reduces the estimate to about 0.25 log point.

**Alternative definitions.** We include two additional robustness exercises. In the first, we define a local labor market within a commuting zone, which typically covers more than one county and might cross state lines. Consistent with our baseline results, low-monopsony-power firms are significantly more responsive to a monetary surprise, although the differential response is even larger, about 0.6

log point four quarters after the shock. Second, the rightmost column of Figure 2a shows the results of our baseline specification, but considering monetary shocks as in Bauer and Swanson (2023) which accounts for the Fed information effect in a different way than Jarociński and Karádi (2020). Our results, however, do not change significantly if one uses this alternative definition of monetary surprises.

**Employment response.** As a final robustness check, we study the relative response of firm employment to a monetary policy shock. Our results, shown in Figure 2b, suggest that the employment response is weaker than that of the wage bill, but still statistically larger for low-monopsony-power firms. In fact, after four quarters, the employment for low-monopsony-power firms is about 0.20 log point higher than that of high-monopsony-power firms, representing about one-third of the overall response of the wage bill. The other two-thirds is accounted for by the differential response of the wage rate, which is significantly higher for low-monopsony-power firms, consistent with the higher passthrough from shocks to wages observed in the empirical literature for small firms within a labor market (Chan et al., 2023).

# 3 Model

Next, we present our oligopsonistic New Keynesian model with heterogeneous firms. To model imperfect competition, we build on the approach pioneered in the trade literature by Atkeson and Burstein (2008) and adapted to labor markets by Berger et al. (2022). Firms are atomistic on national markets but large on their local labor market. This type of model is well-suited conceptualize the empirical evidence we presented in Section 2, because it provides a theory of market shares and passthrough at the firm level.

The distinguishing feature of our model is wage stickiness. As we show, wage stickiness is necessary to replicate the heterogeneous passthrough of monetary policy shocks. However, wage stickiness breaks the block recursivity property that makes most models of this type tractable.[5] Notable exceptions of models with nominal rigidites include Mongey (2021) who focuses on duopolies, and Wang and Werning (2022) who focus on symmetric firms. Relative to these seminal papers, we allow for any degree of firm heterogeneity in terms of productivity and number of local competitors, which enables our model to match the cross-sectional moments that modulate our quantitative results. The key assumption that makes our model tractable is that firms formulate their wage strategies in sequence space and can credibly commit to them. Given this assumption, we can leverage the sequence-space Jacobian method to solve for Nash equilibria on thousands of local markets inhabited by hundreds of thousands of heterogeneous firm, in a way that depends on monetary policy and is consistent with general equilibrium.

[5]For example, Berger et al. (2022) prove that their model is block recursive, meaning that local labor market equilibria can be characterized independently of aggregates. Wage stickiness breaks block recursivity, because it forces firms to play a dynamic game in which aggregate fluctuations are payoff relevant.

**Environment.** Time is discrete and each period corresponds to a quarter. The economy consists of a representative household, a continuum of firms, and a central bank. There are two types of firms: producers and retailers. Producers inhabit a continuum of local labor markets $j \in [0,1]$, each with a finite number of firms indexed by $i \in \{1, 2, \ldots, M_j\}$. They hire labor from the household and produce a homogeneous intermediate good which they sell to retailers on a national market. Retailers differentiate the intermediate good into imperfectly substitutable varieties indexed by $k \in [0,1]$ which they sell to the household. Producers have monopsony power on the labor market. Retailers have monopoly power on the final goods market. We separate producers and retailers, and assume that retailers are not affected by the boundaries of local labor markets, for tractability.

## 3.1 Households

**Setup.** The representative household chooses sequences of consumption $\{c_{kt}\}$ from retailer $k$, labor supply $\{n_{ijt}\}$ to producer $ij$, and nominal bonds $\{B_t\}$ to maximize the present value of expected utility. Given initial wealth $B_0$, the household solves

$$\max_{c_{kt}, n_{ijt}, B_t} \mathbb{E}_0 \sum_{t=0}^{\infty} \beta^t U(C_t, N_t), \tag{3.1}$$

where the aggregate consumption and labor supply indexes are given by

$$C_t = \left( \int_0^1 c_{kt}^{\frac{\epsilon-1}{\epsilon}} dk \right)^{\frac{\epsilon}{\epsilon-1}}, \qquad N_t = \left( \int_0^1 n_{jt}^{\frac{\zeta+1}{\zeta}} dj \right)^{\frac{\zeta}{\zeta+1}}, \qquad n_{jt} = \left( \sum_{i=1}^{M_j} n_{ijt}^{\frac{\eta+1}{\eta}} \right)^{\frac{\eta}{\eta+1}}, \tag{3.2}$$

with elasticities $\epsilon > 1$, and $\eta > \zeta > 0$. The ordering of the labor supply elasticities implies that the representative household is more willing to substitute labor within a local labor market than across markets. Utility maximization is subject to the following budget constraint every period

$$\int_0^1 p_{kt} c_{kt} dk + B_t = (1 + i_{t-1}) B_{t-1} + \int_0^1 \left( \sum_{i=1}^{M_j} W_{ijt} n_{ijt} \right) dj + \Pi_t, \tag{3.3}$$

where $p_{kt}$ is the price of final good $k$, $i_{t-1}$ is the nominal return on bonds between periods $t-1$ and $t$, $W_{ijt}$ is the nominal wage of producer $i$ on local labor market $j$, and $\Pi_t$ is aggregate profits.

**Optimality conditions.** Goods demand is summarized by an aggregate Euler equation and retailer-specific demand curves

$$U_C(C_t, N_t) = \beta \mathbb{E}_t \left[ \frac{1 + i_t}{\pi_{t+1}} U_C(C_{t+1}, N_{t+1}) \right], \qquad c_{kt} = \left( \frac{p_{kt}}{P_t} \right)^{-\epsilon} C_t, \tag{3.4}$$

where the aggregate price index, defined by $P_t C_t = \int p_{kt} c_{kt} dk$, and inflation can be expressed as

$$P_t = \left( \int_0^1 p_{kt}^{1-\epsilon} dk \right)^{\frac{1}{1-\epsilon}}, \qquad \pi_t = \frac{P_t}{P_{t-1}}. \tag{3.5}$$

Similarly, aggregate and firm-specific labor supply are given by

$$\frac{W_t}{P_t} = -\frac{U_N(C_t, N_t)}{U_C(C_t, N_t)}, \qquad n_{ijt} = \left(\frac{W_{ijt}}{W_{jt}}\right)^{\eta} \left(\frac{W_{jt}}{W_t}\right)^{\zeta} N_t, \tag{3.6}$$

where local and aggregate wage indexes are

$$W_t = \left(\int_0^1 W_{jt}^{1+\zeta} dj\right)^{\frac{1}{1+\zeta}}, \qquad W_{jt} = \left(\sum_{i=1}^{M_j} W_{ijt}^{1+\eta}\right)^{\frac{1}{1+\eta}}. \tag{3.7}$$

Note that both firm-specific labor supply and the two wage indexes are linearly homogeneous in wages. Therefore, these equations are valid for both nominal and real wages. Our convention is to use upper case letters to denote nominal variables and lower case letters to denote real variables. See appendix A.1 for derivations.

**Labor market power.** The CES labor supply system described by equations (3.6) and (3.7) implies that, as long as the elasticities $\eta$ and $\zeta$ are finite, producers will have labor market power. Producer $ij$ may offer a lower wage than its local competitors ($W_{ijt} < W_{jt}$) and national competitors ($W_{ijt} < W_t$) and it will still attract workers. Finite labor supply elasticities capture, in reduced form, the deeper reasons for firms' having labor market power: mobility costs, search costs, specialized human capital, and heterogeneous preferences over non-wage amenities (Berger et al., 2022).

### 3.2 Retailers

**Setup.** There is a unit measure of retailers $k \in [0, 1]$ who the buy homogeneous intermediate good for price $M_t$ and turn it into differentiated final goods $c_{kt}$ which they sell to households at price $p_{kt}$. Since varieties are imperfect substitutes, retailers have monopoly power and can set their own price. There is price stickiness à la Calvo: retailers can adjust their price with probability $1 - \theta_p \in (0, 1]$. The probability of adjustment is independently and identically distributed across periods and firms.

A retailer that cannot adjust its price has no decision to make. It produces just enough to satisfy household demand at the current prices. A retailer that can reset its price will maximize the present value of expected profits until the next price adjustment

$$\max_{p_{kt}^*, \{c_{kt+\tau}, y_{kt+\tau}\}} \mathbb{E}_t \sum_{\tau=0}^{\infty} \theta_p^{\tau} R_{t,t+\tau} \left(p_{kt}^* c_{kt+\tau} - M_{t+\tau} c_{kt+\tau}\right) \tag{3.8}$$

subject to the demand curve

$$c_{kt+\tau} = \left(\frac{p_{kt}^*}{P_{t+\tau}}\right)^{-\epsilon} C_{t+\tau}, \tag{3.9}$$

where $R_{t,t+\tau}$ is the nominal discount factor between periods $t$ and $t+\tau$

$$R_{t,t+\tau} = \begin{cases} 1 & \text{for } \tau = 0, \\ \prod_{s=0}^{\tau-1} \frac{1}{1+i_{t+s}} & \text{for } \tau > 0. \end{cases} \tag{3.10}$$

**Price setting equation.** Although the retailer's problem is standard, it is useful to consider it due to its similarity to the novel problem of the oligopsonistic producers. Price adjusters are symmetric and set the same price

$$p_t^* = \frac{\epsilon}{\epsilon - 1} \frac{\mathbb{E}_t \sum_{\tau=0}^{\infty} \theta_p^\tau R_{t,t+\tau} P_{t+\tau}^\epsilon C_{t+\tau} M_{t+\tau}}{\mathbb{E}_t \sum_{\tau=0}^{\infty} \theta_p^\tau R_{t,t+\tau} P_{t+\tau}^\epsilon C_{t+\tau}}. \tag{3.11}$$

Equation (3.11) shows that all retailers have the same market power, captured by a *desired markup* that is constant and equal to $\epsilon/(\epsilon - 1)$. The firm sets $p_t^*$ to achieve this markup over the expected average marginal cost until the next price adjustment. The marginal cost $\tau$ periods ahead is discounted with the nominal discount factor $R_{t,t+\tau}$, with the probability $\theta_p^\tau$ that the price is still in effect, and weighted by expected sales which are proportional to $P_{t+\tau}^\epsilon C_{t+\tau}$.

Loglinearizing the first order condition (3.11) in conjunction with the price index (3.5) yields the textbook New Keynesian Phillips curve

$$\hat{\pi}_t = \frac{(1-\theta_p)(1-\beta\theta_p)}{\theta_p} \hat{m}_t + \beta \hat{\pi}_{t+1}, \tag{3.12}$$

where $\hat{\pi}_t$ and $\hat{m}_t$ denote inflation and real marginal cost in log-deviations from the zero-inflation steady state. See Appendix A.2 for derivations.

### 3.3 Producers

The producers' problem is ultimately similar to the textbook price setting problem of the retailers, with the key difference that it allows for heterogeneous labor market power across firms. There is a conceptual difference as well, namely that firms play a noncooperative dynamic game with their local competitors. We describe the game under full commitment and show that the resulting wage strategies are time consistent.

**Setup.** There is a continuum of producers who are ex-ante heterogeneous in two dimensions. First, each firm operates on a specific local labor market $j \in [0, 1]$ inhabited by finitely many firms $i \in \{1, 2, \ldots, M_j\}$. Thus, each firm faces different local competitors. Second, firms differ in their productivity $z_{ij} \in (0, \infty)$ which is drawn from a location invariant distribution $F(z)$ once and for all. In addition, producers face nominal wage adjustment costs à la Calvo. They can reset their wage with probability $1 - \theta_w \in (0, 1]$ that is independently and identically distributed over time and across firms. Thus, firms also differ ex post in their histories of Calvo shocks $\{\Theta_{ijt}\}_{t=0}^{\infty}$, where $\Theta_{ijt}$ equals one when firm $i$ in sector $j$ can adjust its wage at time $t$, and zero otherwise.

This market structure implies that producers are infinitesimal with respect to the macroeconomy but large on their local labor market. Accordingly, firms take aggregate outcomes as if they were exogenous while engage with their local competitors in a dynamic game to determine local labor market outcomes. Specifically, we assume that firms play Bertrand competition with perfect commitment. Let $s_t$ denote the state of the world in period $t$ and $s^t$ the history of states, $s^t = \{s_0, \ldots, s_t\}$. The state $s_t$ consists of a monetary policy shock $\epsilon_t$ as well as the Calvo shocks for all firms $\{\{\Theta_{ijt}\}_{i=1}^{M_j}\}_{j\in[0,1]}$. In period 0, all firms commit to a (nominal) wage strategy $W_{ij}(s^t)$ that is a best response to the wage strategies of other firms as well as to the aggregate outcomes. As is standard in New Keynesian models, all firms are required to accommodate labor supply at the prevailing wages.

Formally, firm $i$ on market $j$ solves

$$\max_{\{W_{ij}(s^t), n_{ij}(s^t), W_j(s^t)\}} \mathbb{E}_0 \sum_{t=0}^{\infty} \sum_{s^t} R(s^t) \left[ M(s^t) z_{ij} n_{ij}(s^t) - n_{ij}(s^t) W_{ij}(s^t) \right], \tag{3.13}$$

$$\text{s.t.} \qquad n_{ij}(s^t) = \left( \frac{W_{ij}(s^t)}{W_j(s^t)} \right)^{\eta} \left( \frac{W_j(s^t)}{W(s^t)} \right)^{\zeta} N(s^t), \tag{3.14}$$

$$W_j(s^t) = \left[ W_{ij}(s^t)^{1+\eta} + \sum_{i' \neq i} W_{i'j}(s^t)^{1+\eta} \right]^{\frac{1}{1+\eta}}, \tag{3.15}$$

$$W_{ij}(s^t) = W_{ij}(s^{t-1}) \quad \text{if } \Theta_{ijt} = 0. \tag{3.16}$$

The firm sets its on wage $W_{ij}(s^t)$ to maximize the present discounted value of future profits, taking into account the labor supply curve (3.14). In doing so, it considers its ability to move the local wage index $W_j(s^t)$, given the age strategies of its local competitors $W_{i'j}(s^t)$, as shown by (3.15). Finally, (3.16) ensures that the wage strategy is implementable given the Calvo shocks.

**Separability across wage spells.** Next, we argue that the optimal wage strategy may be solved for in a piece-wise fashion, considering wage spells in isolation.

Definition **1.** A wage spell $\mathcal{S}_{ijt}$ is a set of states $s^{t+\tau}$ with $\tau \geq 0$ such that: (i) the history of states is equivalent up to time $t$, (ii) the firm gets to adjust its wage at time $t$, $\Theta_{ijt} = 1$, (iii) the firm is stuck with its time-$t$ wage until time $\tau$, that is $\tau$ is the largest number such that $\sum_{\tilde{\tau}=1}^{\tau} \Theta_{ijt+\tilde{\tau}} = 0$.

Given this definition, we can partition the firm's problem into wage spells and establish separability between wage spells, allowing us to write the wage setting problem as

$$\max_{\left\{W_{ijt}^*, n_{ijt+\tau}, W_{jt+\tau}\right\}} \mathbb{E}_{\mathcal{S}_{ijt}} \sum_{\tau=0}^{\infty} R_{t+\tau} \left[ M_{t+\tau} z_{ij} n_{ijt+\tau} - n_{ijt+\tau} W_{ijt}^* \right], \tag{3.17}$$

$$\text{s.t.} \qquad n_{ijt+\tau} = \left( \frac{W_{ijt}^*}{W_{jt+\tau}} \right)^{\eta} \left( \frac{W_{jt+\tau}}{W_{t+\tau}} \right)^{\zeta} N_{t+\tau}, \tag{3.18}$$

$$W_{jt+\tau} = \left[ \left( W_{ijt}^* \right)^{1+\eta} + \sum_{i' \neq i} W_{i'jt+\tau}^{1+\eta} \right]^{\frac{1}{1+\eta}}. \tag{3.19}$$

The derivation, relegated to Appendix A.3, hinges on our assumption of full commitment. Without commitment, firms would have to consider the off-equilibrium effects that their own wage choice today has on their competitors future wage choices, leading to dependence between wage spells.[6] At the notational level, the only difference between the objective functions of retailers and producers is that retailers only care about their own Calvo shocks and hence the expectations in (3.8) are straightforward to write out without resorting to "price spells".

**Wage setting equation.** In Appendix A.3, we derive the optimal reset wage

$$W^*_{ijt} = \frac{\mathbb{E}_{\mathcal{S}_{ijt}} \sum_{\tau=0}^{\infty} R_{t,t+\tau} n_{ijt+\tau} \left[\eta - (\eta - \zeta) s_{ijt+\tau}\right] M_{t+\tau} z_{ij}}{\mathbb{E}_{\mathcal{S}_{ijt}} \sum_{\tau=0}^{\infty} R_{t,t+\tau} n_{ijt+\tau} \left[1 + \eta - (\eta - \zeta) s_{ijt+\tau}\right]}, \tag{3.20}$$

where $s_{ijt} \equiv W_{ijt} n_{ijt} / (W_{jt} n_{jt})$ is the wage bill share of firm $ij$ on market $j$ in period $t$. Compare this expression to the well-known price setting equation of retailers (3.11). Firm $ij$ sets $W^*_{ijt}$ to implement a *desired markdown* with respect to the marginal revenue product of labor (MRPL), $M_{t+\tau} z_{ij}$. It considers the expected present value of MRPL until the next wage adjustment, discounted and weighted by employment. The crucial difference from monopolistic competition is that the desired markdown

$$\mu_{ijt} \equiv \frac{\epsilon_{ijt}}{1 + \epsilon_{ijt}} \equiv \frac{\eta - (\eta - \zeta) s_{ijt}}{1 + \eta - (\eta - \zeta) s_{ijt}} \tag{3.21}$$

may vary both across firms and over time. Given the ordering $\eta > \zeta > 0$, firms with larger market share $s_{ijt}$ set lower wages relative to MRPL. In the context of our model, firm heterogeneity in market shares originates in differences in productivity and in the number of competitors. Equation (3.21) highlights that we can interpret the markdown in terms of an equilibrium *labor supply elasticity*, $\epsilon_{ijt}$, which summarizes the market power of firm $ij$ both within its local labor market and on the national labor market. Equation (3.21) also clarifies that the relevant measure of labor market power for the purposes of wage setting (and thus for determining labor demand and production) is the wage bill share. This result guides our classification of firms by local wagebill share in our empirical analysis.

**Flexible-wage limit.** The mapping (3.21) from the primitive labor supply elasticities $\eta$ and $\zeta$ to the desired markdown $\mu_{ijt}$ is the same as in the flexible-wage limit of the model. That is the model studied in Berger et al. (2022), with the one difference that they focus on Cournot competition.[7] We assume Bertrand competition, because formulating strategies in terms of wages is more natural than in terms of employment when wages are sticky. Furthermore, the equilibrium of the flexible-wage limit is the same as the deterministic steady state of our sticky-wage model. In sum, in response to small monetary policy shocks, the distributions of wages, employment, and markdowns fluctuate around the Nash equilibrium of the static benchmark model of oligopsony.

[6] We discuss the assumption of commitment in more detail in Section 3.6.

[7] Cournot competition implies that the equilibrium elasticity is a harmonic mean

$$\epsilon^{Cournot}_{ijt} = \left[\frac{1}{\eta} - \left(\frac{1}{\eta} - \frac{1}{\zeta}\right) s_{ijt}\right]^{-1}.$$

### 3.4 Central Bank

The central bank sets the nominal rate according to a standard Taylor rule

$$i_t = i_{ss} + \phi_\pi(\pi_t - 1) + \varepsilon_t. \tag{3.22}$$

The ex-post real return on nominal bonds satisfies the Fisher equation

$$1 + r_t = \frac{1 + i_{t-1}}{1 + \pi_t}. \tag{3.23}$$

### 3.5 Equilibrium

Given a sequence of monetary policy shocks $\{\varepsilon_t\}$, sequences of Calvo shocks for each producer $\{\Theta_{ijt}\}$, and initial wealth $B_{-1}$, the equilibrium is a sequence of allocations $\{B_t, C_t, N_t, n_{jt}, n_{ijt}\}$ and prices $\{i_t, r_t, \pi_t, w^*_{ijt}, w_{ijt}, w_{jt}, w_t, m_t\}$ such that

1. Households and producers optimize;
2. Local labor markets are in Nash equilibrium with real wages and wage indexes[8] satisfying (3.7);
3. Inflation satisfies the loglinearized Phillips curve[9] (3.12);
4. The nominal rate satisfies the Taylor rule and the real rate satisfies the Fisher equation;
5. Goods and bond markets clear

$$Y_t \equiv \int_0^1 \sum_{i=1}^{M_j} z_{ij} n_{ijt} dj = C_t, \qquad B_t = 0. \tag{3.24}$$

### 3.6 Discussion

Before moving on, let us discuss some of the key assumptions of our model.

First, our equilibrium concept assumes that firms have full commitment when setting their wage-path strategies. This assumption enables us to solve the model at the scale that is required to match the empirical evidence. An alternative equilibrium concept which we do not consider is the Markov perfect equilibrium (MPE). Reinganum and Stokey (1985) note that the two approaches correspond to two extreme assumptions about players' ability to commit to future actions. Using path strategies, like we do in this paper, corresponds to the assumption that firms can commit for their entire planning horizon. Using decision rule strategies, as in Mongey (2021) and Wang and Werning (2022), corresponds to the assumption that no commitment is possible at all.[10]

A key difference between our equilibrium concept and the MPE is in the level of markdowns in the non-stochastic steady state. In our setup, those would be identical to the flexible-wage equilibrium.

[8] The wage indexes are stated in terms on nominal wages, but they apply to real wages $w_t \equiv W_t/P_t$ as well.

[9] Working directly with the loglinearized Phillips curve lets us omit the second-order effects of price dispersion.

[10] We note that in our setting, the full-commitment equilibrium does not lead to time-inconsistent solutions as is often the case. This is because a firm is atomistic with respect to the aggregate economy and it takes the wage paths of its competitors as given.

With MPE, steady-state markdowns will be higher than in the flexible-wage equilibrium. This is because a firm internalizes how raising its markup in the current period can lead to a future increase in the markdowns of its competitors.[11] Since the focus of this paper is the transmission of monetary policy with oligopsinstic competition, the concern regarding the choice of equilibrium is not about the steady-state level but rather the response of the economy to monetary policy. This concern is attenuated by the finding of Wang and Werning (2022), who study the transmission of monetary policy with oligopolistic competition. They show that while the steady-state allocations under MPE and under full-commitment are significantly different, the difference in the response of the economy to a monetary policy shock is miniscule.[12]

Second, our model assumes that each local labor market is populated by a finite number of oligopsonistic firms, and that there is constant elasticity of substitution between firms within the local labor market. As we show in the next section, the qualitative ability of the model to match the heterogeneous response observed in the data hinges on optimal markdowns being an increasing and convex function of the relative market share of the firm. Such markdown behavior arises with a nested-CES structure with a oligopsonistic competition in the inner nest akin to Atkeson and Burstein (2008) and Berger et al. (2022), as we adopt in this paper. Alternatively, we could have assumed that there is a continuum of firms within each local labor market but that the elasticity of substitution across these firms is not constant, e.g., Kimball (1995). As long as optimal markdowns remain increasing and convex in the relative market share of the firm, such change will not affect the qualitative response of the economy to a monetary policy shock. We opted for the nested-CES with oligopsonistic competition structure as it provides a cleaner mapping between model and data. This is because it allows us to: (i) classify firms in the model to low- and high-labor-market-power groups in the same way as we do in the data, (ii) match the degree of HHI observed in the data.

Third, our model assumes that both wages and prices are subject to nominal rigidities. Having two sources of nominal rigidities may seem redundant, but that is not the case. Each source of nominal rigidity serves a different purpose in the ability of the model to match key features of the data. First, as we show in the next section, nominal wage rigidities are necessary to obtain the heterogeneous response of firms to monetary policy observed in the data. Second, nominal price rigidities are necessary to match the *average* wage bill response. If prices were flexible, an expansionary demand shock would cause both real wages and employment to fall. With nominal price rigidities, an expansionary monetary policy can lead to an increase in both wages and employment.

[11] Competitors' future markdowns may rise for two reasons. First, if the firm is still stuck at its current wage while a competitor resets theirs, strategic complementarities imply that the competitor will optimally choose a higher markdown. Second, even after the firm resets its own wage, that wage can continue to influence future markdowns—through competitors who adjusted their wage during the firm's previous wage spell and remain locked into that level.

[12] Wang and Werning (2022)'s measure of monetary non-neutrality is the half-life of the price-level response to a permanent monetary shock. They report that the difference in the half-life under MPE and full commitment is less than 3 percent in all of their specifications. To put that in context, moving from monopolistic to oligopolistic competition increases the half-life by 40 percent in their framework.

# 4 Monetary Policy Transmission with Oligopsonistic Competition

We next study the transmission of monetary policy in our model. To make the analysis transparent, we consider an analytically tractable special case and characterize the behavior of a local labor market in partial equilibrium. This exercise serves three purposes. First, it shows that nominal wage rigidities are necessary for the model to generate the heterogeneous effects of monetary policy across low- and high-labor-market-power firms observed in the data. Second, it illustrates how the nested-CES structure with oligopsonistic competition naturally gives rise to greater responsiveness among low-labor-market-power firms. Third, it highlights the two channels through which oligopsonistic competition dampens the aggregate effect of monetary policy on output: (i) partial pass-through and (ii) misallocation.

We focus on a local labor market with three firms that differ in productivity levels, $z_1 < z_2 < z_3$. We assume myopic households ($\beta = 0$), so the wage-setting decision is static. Under these assumptions, the optimal wage-setting condition simplifies to

$$W_{it}^* = \frac{1}{\mu_{it}} z_i M_t, \tag{4.1}$$

where $M_t$ is the price of the good sold by the oligopsonistic firms, which they take as given. $\mu_{it}$ is the desired markdown of firm $i$ at time $t$, which is given by

$$\mu_{it} = \frac{1 + \eta - (\eta - \zeta) s_{it}}{\eta - (\eta - \zeta) s_{it}}, \tag{4.2}$$

with $s_{it}$ denoting the firm's wage-bill share in the local labor market. In the full model, the transmission of monetary policy to employment and wages depends on how each local market responds to a change in $M_t$. Accordingly, we analyze the response of a local labor market to an increase in $M_t$.

Total differentiation of the wage-setting equation yields

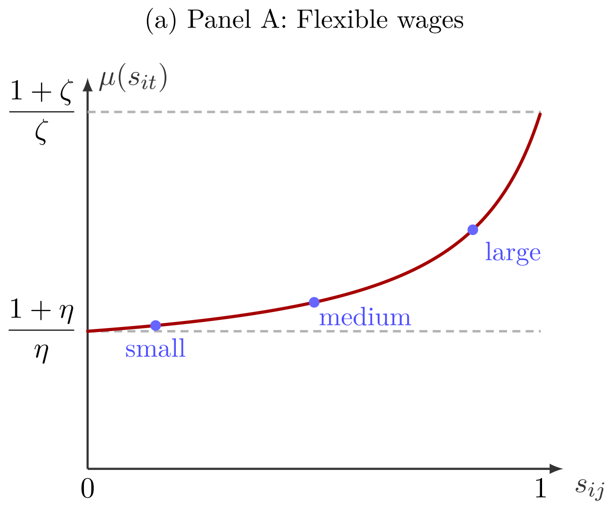
$$\frac{\partial \ln W_{it}^*}{\partial \ln M_t} = 1 - \frac{\partial \ln \mu_{it}}{\partial \ln M_t}. \tag{4.3}$$

Absent a change in the optimal markdown, the elasticity of wage response to the price of goods sold is one. If the markdown optimally rises in response to an increase in the price, there is only partial passthrough: wages adjust less than one-for-one. As equation (4.2) shows, the optimal markdown depends on each firm's wage-bill share, which in turn depends not only on its own wage but also on the wages of competitors in the local labor market. To understand this relationship, we next solve for the equilibrium in the local labor market and highlight the key ingredients that allow the model to replicate the heterogeneous responses across firms observed in the micro data.

## 4.1 The Heterogeneous Passthrough of Monetary Policy

We begin by analyzing the equilibrium in the local labor market when wages are fully flexible. Consider a rise in the intermediate input price at time $t = 1$. Let $\Delta \ln x_i$ denote the log difference between any variable $x$ of firm $i$ in period 1 and its initial level at time 0. The following proposition characterizes the resulting local labor market equilibrium.

Figure 3: Heterogeneous Response to Monetary Policy in Special Case




**Notes**: The panels illustrate how markdowns and wage-bill shares respond to an increase in the intermediate good's price under two scenarios. In Panel A (flexible wages), all firms fully passthrough the price increase, leaving wage-bill shares and markdowns unchanged. In Panel B (rigid wage), the *medium* firm's nominal wage is fixed, causing its markdown to rise. In equilibrium, the small and large firms expand employment, but because the markdown function is convex, the large firm's markdown rises more while its wage-bill share increases less than that of the small firm.

Proposition **1.** *Consider an increase in the intermediate input price at $t = 1$, so that $M_1 > M_0$. If all firms can adjust their wages, there is a homogeneous response of wage and employment across firms. That is, $\Delta \ln W_i = \Delta \ln W_{i'}$ for all $i$ and $i'$, and $\Delta \ln s_i = 0$ for all $i$.*

All formal proofs are provided in Appendix A.4. Proposition 1 implies that, in the absence of wage rigidities, both wage-bill and employment responses are homogeneous across firms. This homogeneity arises despite differences in steady-state markdowns across firms.

Panel A of Figure 3 illustrates the intuition behind the homogeneous passthrough result. The red line plots the optimal markdown as a function of the wage-bill share. Suppose all firms fully passthrough the rise in the good's price to their wages. In that case, all wages rise proportionally, and each firm's wage-bill share at time $t = 1$ remains identical to its share at $t = 0$. Because wage-bill shares are unchanged, so are the desired markdowns. Hence, following equation (4.3), our conjecture that full passthrough is the equilibrium outcome is confirmed.

Two key modeling assumptions underlie the homogeneous passthrough result. First, in our framework, the wage-bill share is a sufficient statistic that fully characterizes the optimal markdown. Second, wage-bill shares are homogeneous of degree zero. These assumptions are not specific to the nested-CES preferences we use for the disutility of labor and also hold under alternative formulations such as Kimball (1995) preferences.

We now turn to the equilibrium with nominal wage rigidities. Suppose that one of the three firms cannot adjust its nominal wage. The following proposition shows that in this case, the response becomes heterogeneous across the two firms that can adjust their wages. In particular, both the wage and employment of the smaller firm respond more strongly.

Proposition **2.** *Suppose that firm $i$ cannot adjust its wage, so that $\Delta \ln w_i = 0$. Then, among the firms that can adjust their wages, the wage and employment responses are larger for the smaller of the* *two. That is,* $\Delta \ln s_j < \Delta \ln s_{j'}$ *for* $j, j' \neq i$ *and* $z_j > z_{j'}$.

To build intuition for Proposition 2, suppose that the medium firm's nominal wage is rigid. Because its wage remains fixed when the price of the good rises, its markdown increases, shifting employment toward the flexible firms. The key question is whether the flexible firms adjust their wages symmetrically. The answer depends on the curvature of the markdown function in equation (4.2), which is governed by the superelasticity of labor supply. When the superelasticity decreases with the wage-bill share, the markdown function is convex, implying that larger firms face a flatter residual labor supply than smaller firms.

Panel B of Figure 3 illustrates this mechanism. When the medium firm's wage is rigid, both the small and large firms expand employment. Because the markdown function is convex, an equal proportional increase in wage-bill shares induces a larger rise in the desired markdown for the relatively larger firm. The smaller firm therefore raises its wage relatively more, while the larger firm's passthrough is dampened. The smaller firm's stronger wage response also translates into a larger increase in employment. The heterogeneous transmission of monetary policy consistent with our micro evidence thus arises whenever the superelasticity of labor supply decreases with firm size—a condition that holds under the nested-CES structure with $\eta > \zeta$ that we assume.

## 4.2 Oligopsonistic Competition and the Effectiveness of Monetary Policy

After establishing the mechanism underlying the heterogeneous response across firms, we now study the aggregate effect of oligopsonistic competition on the transmission of monetary policy. There are two channels through which oligopsonistic competition dampens the response of employment and output to an expansionary monetary shock: a partial passthrough channel and a misallocation channel.

The first channel operates through incomplete passthrough of prices to wages. With nominal rigidities, the passthrough from the price of goods sold, $M$, to wages is below one. Part of this limited adjustment is mechanical—firms with fixed wages exhibit zero passthrough—but this component is independent of the degree of competition. Oligopsonistic competition further reduces passthrough because firms able to adjust their wages choose to raise their markdowns rather than fully transmit higher prices to wages. When average markdowns rise, the passthrough from prices to wages is incomplete, so nominal shocks translate less into real wage adjustments and employment growth. Intuitively, oligopsonistic firms absorb part of the monetary expansion by widening their markdowns, weakening the overall sensitivity of labor demand to monetary policy.

The degree of competition in the market directly shapes the extent of passthrough. As shown in Figure 3, the markdown is a convex function of the wage-bill share, implying that passthrough from prices to wages declines with firm size. Consequently, as the number of firms in a local labor market increases, the average passthrough rises, reducing the strength of the partial-passthrough channel.

The second channel through which oligopsonistic competition dampens the effect of monetary policy is the misallocation channel. Recall that goods produced by the oligopsonistic firms are perfect substitutes. To a first-order approximation, aggregate production in the economy is given by[13]

[13]The relationship holds only to a first order, because nominal price rigidities introduce a distortion (Yun, 2005).

$$Y_t \approx \int_0^1 \sum_i^{M_j} z_{ij} n_{ijt} dj. \tag{4.4}$$

Letting $\widetilde{N}_t \equiv \int_0^1 \left(\sum_i n_{ijt}\right) dj$ denote the overall amount of labor supplied, we can define aggregate TFP as

$$A_t \equiv \int_0^1 \left( \sum_i^{M_j} \frac{n_{ijt}}{\widetilde{N}_t} z_{ij} \right) dj, \tag{4.5}$$

so that $Y_t \approx A_t \widetilde{N}_t$. Equation (4.5) implies that a shift of employment from high- to low-productivity firms reduces aggregate TFP.[14] Since changes in the wage-bill share mirror changes in employment, a decline in the wage-bill share of high-productivity firms leads to a reduction in aggregate TFP.

Proposition 2 reveals the direction of the misallocation channel. In the presence of nominal rigidities, a rise in demand increases the wage-bill share of low-productivity firms and reduces that of high-productivity firms, leading to a decline in aggregate TFP. Two observations are worth emphasizing. First, misallocation arises only in the presence of heterogeneous firms. Firm heterogeneity is therefore not merely a modeling feature aimed at matching the micro evidence but a key element shaping the transmission of monetary policy under oligopsonistic competition. Second, our main finding is not that expansionary monetary policy reduces TFP, but that oligopsonistic competition dampens the effect of monetary policy through its effect on TFP. The overall response of TFP to monetary shocks may also depend on other mechanisms—such as changes in capital utilization—that are outside our framework and orthogonal to oligopsonistic competition.

The two channels we highlight—partial passthrough and missalocation—are not entirely novel, and have been studied in some form in the context of imperfect competition in the product market. The partial-passthrough channel is present in Mongey (2021) and Wang and Werning (2022), who study monetary transmission under oligopolistic competition among homogeneous firms. The misallocation channel is present Baqaee et al. (2024), who analyze monetary policy when firms differ in both market power and price passthrough. One might therefore conjecture that imperfect competition has similar implications for monetary transmission regardless of whether it arises in product or labor markets. Our analysis shows that the opposite is true.

Whether in the product market or in the labor market, imperfect competition generates partial and heterogeneous passthrough across firms. The difference in its implications for monetary policy stems from whether these passthroughs affect the supply or demand side of the market. Imperfect competition in the labor market—the focus of this paper—produces partial and heterogeneous passthrough in the demand for labor. Because households are on their labor-supply curve, partial passthrough to wages leads to a smaller employment response. In contrast, imperfect competition in the product market leads to partial passthrough to prices, which results in a *larger*, not smaller, quantity response. Consequently, Mongey (2021) and Wang and Werning (2022) find the oligopolistic competition among homogeneous firms amplifies the output response to monetary policy. And Baqaee et al. (2024) find that lower passthrough by high-markup firms allows expansionary monetary policy to raise aggregate

[14] The marginal product of labor is not equalized across firms in the non-stochastic steady state because the marginal disutility of labor is heterogeneous across firms—an implication of the nested-CES structure for the disutility of labor.

TFP by alleviating misallocation.

Taken together, the stylized framework clarifies the core mechanisms through which oligopsonistic competition shapes the transmission of monetary policy. It highlights how nominal wage rigidities and convex markdowns jointly generate heterogeneous firm-level responses and aggregate dampening. We next turn to the full model to quantify the strength of these channels and assess their implications for aggregate employment and output dynamics.

# 5 Quantitative Analysis

This section quantifies how oligopsonistic competition shapes the transmission of monetary policy. We begin by describing the numerical method we develop to solve the model, and then calibrate the model to match key features of the U.S. labor market. The quantitative results show that oligopsonistic competition substantially dampens the real effects of monetary policy: output responds about one-quarter less than in an economy where firms maintain stable markdowns, with roughly two-thirds of the attenuation due to partial passthrough and one-third to misallocation. We conclude by evaluating how the observed decline in local labor market concentration since 1980 has amplified the effectiveness of monetary policy.

## 5.1 Numerical Solution

Most models of market power in the Atkeson and Burstein (2008) tradition are tractable because they involve a static game. Nominal rigidities introduce dynamics. A firm setting its wage has to consider future developments in both aggregate and local conditions, because it may not be able respond to them when they are realized. This subsection discusses the challenges that arise due to the introduction of nominal rigidities and our approach to overcome them.

**Challenges.** There are two main challenges for computing the equilibrium of the model. First, the law of large numbers does not apply to local labor markets with finitely many firms. Second, firms and workers interact in a large number of markets.

To appreciate the challenge posed by idiosyncratic uncertainty on local markets, consider the problem of finding the local equilibrium in a single market. Equation (3.20) shows that the optimal wage of firm $i$ in market $j$ at time $t$ depends on expectations of aggregate variables $\{\mathbb{E}_{\mathcal{S}_{ijt}}[R_{t,t+\tau}], \mathbb{E}_{\mathcal{S}_{ijt}}[M_{t+\tau}]\}_{\tau=0}^{\infty}$ as well as the expectations of the local wage index $\{\mathbb{E}_{\mathcal{S}_{ijt}}[W_{jt+\tau}]\}_{\tau=0}^{\infty}$. Recall that the local wage index is given by

$$W_{jt+\tau} = \left(\sum_{i=1}^{M_j} W_{ijt+\tau}^{1+\eta}\right)^{\frac{1}{1+\eta}}, \qquad W_{ijt+\tau} = \begin{cases} W^*_{ijt+\tau} & \text{if } \Theta_{ijt+\tau} = 1, \\ W_{ijt+\tau-1} & \text{if } \Theta_{ijt+\tau} = 0. \end{cases} \tag{5.1}$$

Since $M_j$ is finite, the expected path of the local wage index until firm $i$ can reoptimize its wage depends on the Calvo shocks of every other firm in market $j$. For concreteness, let us assume that there are 16 firms in the market, the median in our calibrated model, and assume that the probability that a wage spell lasts longer than 20 quarters is negligible. The result is $2^{300} \approx 2 \cdot 10^{90}$ possible permutations of wage adjustments. Evaluating this expectation exactly is impossible even for a single firm, not to mention finding the mutual best responses of 16 firms. And this thought experiment is just about local equilibrium in a single market.

The challenge of working with a large number of markets is that equilibrium prices are characterized implicitly, as solutions to a fixed point problem. Having many markets means having to deal with a large-scale fixed point problem. Our calibrated model features about $975,000$ firms, inhabiting $25,000$ markets. Solving for impulse responses over 100 quarters means characterizing $100 \times 975,000$ firm-level wages, $100 \times 25,000$ local wages plus $100 \times 2$ aggregate prices. This challenge is not present in typical HANK models, in which a distribution affects aggregate demand or supply but agents still interact on a small number of national markets. We have to both keep track of a high-dimensional distribution and solve for many equilibrium prices.

**Solution.** Since evaluating the expectations $\mathbb{E}_{\mathcal{S}_{ijt}}[W_{jt+\tau}]$ under uncertainty about Calvo shocks is infeasible, we endow firms with perfect foresight of both idiosyncratic as well as aggregate shocks.[15] Firms learn the entire sequences of monetary policy shocks and Calvo shocks in period zero, and can compute the resulting paths of endogenous variables. All derivations in Section 3 and Appendix A.3 remain valid. We may simply drop the expectation operator $\mathbb{E}_{\mathcal{S}_{ijt}}$.

The modularity of the SSJ method enables us to overcome the challenge of many markets. The key property of the model we exploit is that firms in market $j$ do not compete directly with firms in market $j'$. This can be seen clearly from the labor supply curve

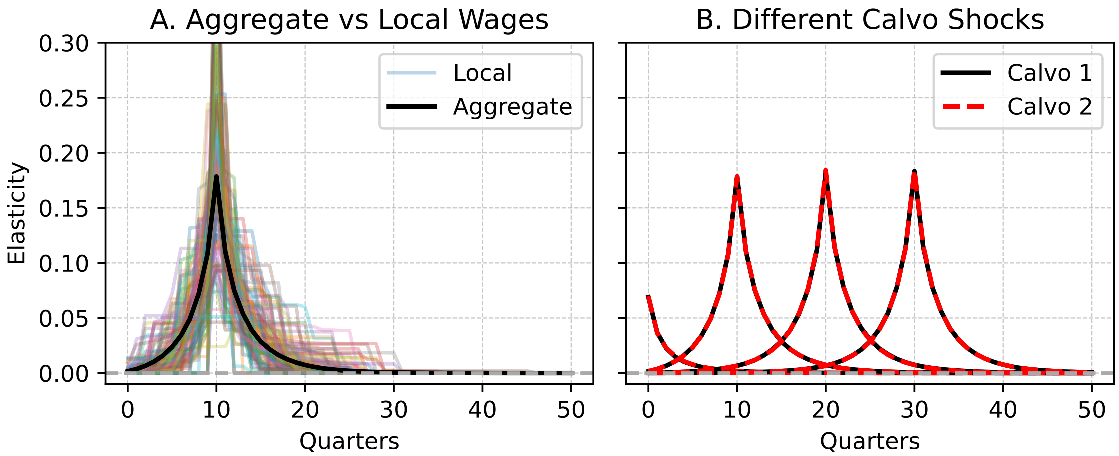
$$n_{ijt} = \left(\frac{W_{ijt}}{W_{jt}}\right)^{\eta} \left(\frac{W_{jt}}{W_t}\right)^{\zeta} N_t. \tag{5.2}$$

Firm $ij$ considers how its wage $W_{ijt}$ compares to the local wage $W_{jt}$ and to the aggregate wage $W_t$. Local wages in other markets matter only through their effect on the aggregate wage. Given this insight, we solve the model in two steps. In the inner layer, we solve for Nash equilibria in local labor markets independently of each other and conditional on aggregates. In the outer layer, we solve for general equilibrium in national markets.

**Firm block.** We use a large but finite number of markets, $J$, and assume that the economy returns to steady state after some finite number of periods $T$. The inner layer of the production block is a collection of $J$ functions mapping $\{w_{jt}, \pi_t, m_t\}_{t=0}^{T}$ into $\{\tilde{w}_{jt}, A_{jt}\}_{t=0}^{T}$. The inputs are the payoff-relevant prices that firms take as given. The outputs are the local wages and productivity that result from firms' best-response. Nash equilibrium corresponds to the fixed point $\tilde{w}_{jt} = w_{jt}$ for all $t$. In SSJ terminology, each local market is a "solved block", $\{\pi_t, m_t\}_{t=0}^{T} \mapsto \{\tilde{w}_{jt}, A_{jt}\}_{t=0}^{T}$, whose Jacobians may be obtained by applying the implicit function theorem and the chain rule as usual. The last step is to aggregate the local solved Jacobians to obtain the Jacobians of the firm block as a whole, which maps

[15] Assuming perfect foresight with respect to aggregate shocks is common in macroeconomics, especially in heterogeneous-agent models. Assuming perfect foresight with respect to idiosyncratic shocks is not common, because it is typically not needed for tractability.

Figure 4: Firm Block Jacobians: Wage with respect to Sales Price



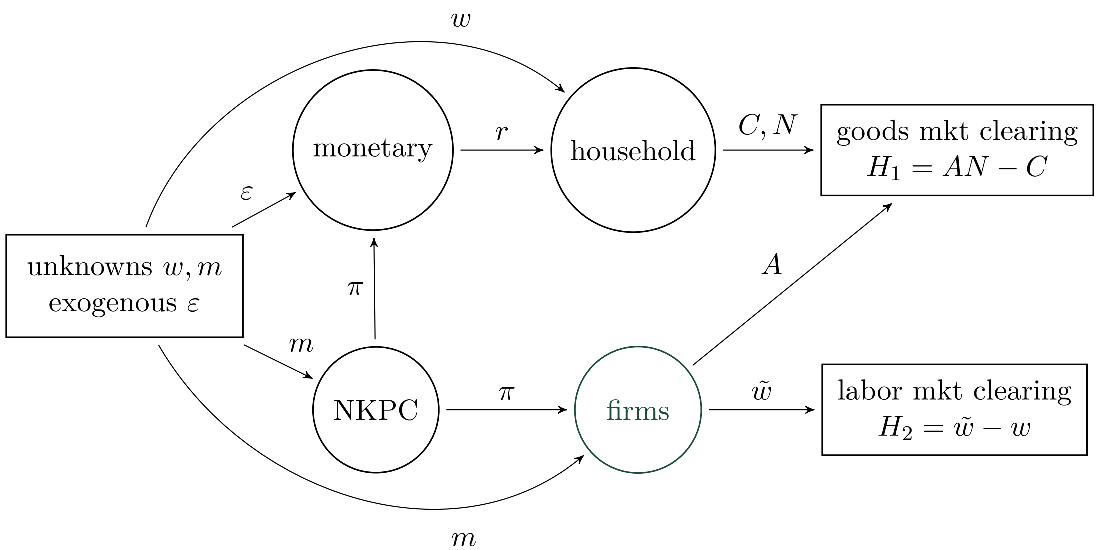

**Note**: Panel A shows column 10 of the Jacobian of the local wage with respect to the sales price for a subsample of local labor markets (in color) as well as the corresponding response of the aggregate wage (in black). Panel B shows columns 0, 10, 20, and 30 of the Jacobian of the aggregate wage with respect to the sales price for two different sets of Calvo shocks.

aggregate prices $\{\pi_t, m_t\}_{t=0}^{T}$ into the aggregate wage and TFP $\{\tilde{w}_t, A_t\}_{t=0}^{T}$. In Appendix B, we walk through each of these steps in detail.

The key feature of our approach is to characterize the Jacobians of the firm block by manipulating many $T \times T$ matrices instead of a few $JT \times JT$ matrices, which would be infeasible for large $J$. We find that a large number of markets, $J = 25,000$, is required to ensure that impulse responses are not sensitive to further increasing the number of markets. Figure 4 illustrates the role of many markets. Panel A compares the Jacobians of local wages $\tilde{w}_{jt}$ for a random subsample of markets with the Jacobian of the aggregate wage $\tilde{w}_t$ with respect to $m_{10}$, a rise in the sales price 10 quarters ahead. The responses of local wages are all different due to heterogeneity in the wage bill shares and Calvo shocks of local firms. Furthermore, the local wage responses display visible steps that correspond to wage adjustments by a single large firm (or simultaneous adjustments by a few firms). In contrast, the Jacobian of the aggregate wage is smooth, because aggregating over sufficiently many markets washes out the impact of individual Calvo shocks. Panel B makes this point more directly, by showing that the Jacobians of the aggregate wage for two different sets of Calvo shocks are visually identical.

**Full model.** With the firm block in hand, we can represent the economy as a directed acyclic graph (DAG) shown by Figure 5. The DAG is a useful tool to conceptualize the monetary transmission mechanism as well as to solve the model numerically. Following the arrows reveals how the monetary policy shock $\varepsilon$ propagates through the economy. The first-round effect of monetary easing—holding prices $w$ and $m$ fixed—is to lower the real interest rate $r$ and stimulate consumption $C$. To meet the increase in aggregate demand, firms have to increase employment $N$ by offering higher real wages $w$. The incentive to do so is provided by a rise in the relative price of intermediate goods $m_t$, which in turn raises inflation $\pi$ and prompts the central bank to raise the nominal rate and curb the original stimulus. From a macroeconomic perspective, the oligopsonistic competition between heterogeneous firms simply

Figure 5: DAG Representation of General Equilibrium




**Note**: Figure shows the dynamic acyclic graph representation of our New Keynesian model. The green firm block is constructed by aggregating a large number of solved blocks, one for each local labor market. The Jacobians of the firm block capture the passthrough of shocks to the real sales price $m_t$ and to inflation $\pi_t$ into the aggregate real wage $\tilde{w}_t$ and aggregate TFP $A_t$.

modulates the increase in $m$ required for aggregate supply to meet the increase in aggregate demand.

## 5.2 Calibration

We distinguish two sets of parameters. First, parameters that are common in New Keynesian models, which we set to conventional values. Second, parameters of the oligopsonistic firm block, which we calibrate internally to match moments of the distribution of firms and local labor markets. Table 1 summarizes the calibration. Table 2 shows the targeted moments which the model fits exactly.

**Fixed parameters.** We assume that the representative household has separable preferences over consumption and labor aggregates

$$U(C_t, N_t) = \frac{C_t^{1-\gamma}}{1-\gamma} - \varphi \frac{N_t^{1+\nu}}{1+\nu}, \tag{5.3}$$

and set the elasticity of intertemporal substitution to $1/\gamma = 0.5$, and the Frisch elasticity to $1/\nu = 0.5$. We normalize aggregate labor supply to $N = 1$ and back out the labor disutility parameter $\varphi$ that justifies this as an equilibrium outcome. We set the quarterly discount factor to 0.99, implying an annual real interest rate of about 4 percent. Steady-state inflation is normalized to zero. The Taylor rule coefficient on inflation is $\phi_\pi = 1.5$ and assume that the monetary policy shock follows and AR(1) process with quarterly autocorrelation $\rho_\varepsilon = 0.85$. We set the elasticity of substitution between retail goods to $\epsilon = 7$, whose only role is to pin down the relative price of intermediate goods in steady state

Table 1: Summary of parameters

| Parameter | | Value | Parameter | | Value |
|---|---|---|---|---|---|
| *A. Fixed parameters* | | | *B. Internally calibrated parameters* | | |
| $\beta$ | Discount factor | 0.99 | $\zeta$ | Across market elasticity | 0.38 |
| $1/\gamma$ | EIS | 0.50 | $\eta$ | Within market elasticity | 2.50 |
| $1/\nu$ | Frisch elasticity | 0.50 | $\xi_z$ | Productivity distribution shape | 3.12 |
| $\epsilon$ | Retail goods elasticity | 7.00 | $\xi_m$ | Market size distribution shape | 0.07 |
| $\theta_p$ | Price stickiness | 0.90 | | | |
| $\theta_w$ | Wage stickiness | 0.80 | | | |
| $\phi_\pi$ | Taylor rule coefficient | 1.50 | | | |
| $\rho_\epsilon$ | Persistence of MP shock | 0.85 | | | |

at $m = (\epsilon - 1)/\epsilon$. We set prices stickiness to $\theta_p = 0.9$, within the range of typical estimates of price Phillips curves without indexation (see e.g. Christiano et al. 2005; Auclert et al. 2020). Finally, we set wage stickiness to $\theta_w = 0.8$, consistent with the findings of Grigsby et al. (2021) who report that 80.6% of job stayers experience no wage change in a given quarter, and Hazell and Taska (forthcoming), who report that posted wages change every 5 quarters on average.

**Internally calibrated parameters.** The remaining parameters are the labor supply elasticities $\eta$ and $\zeta$ as well as the parameters of the distributions of productivity $z$ and market size $M_j$. We assume that the latter are independent Pareto distributions with shape parameters $\xi_z$ and $\xi_m$, which leaves us with four parameters in total. We calibrate these four parameters by targeting an equal number of cross-sectional moments: the average markdown, the employment share of high-monopsony firms, the population share of high-monopsony firms, and mean HHI across local markets. Table 2 Panel A shows the values of these four moments in our baseline calibration, along with the parameters that they identify.[16] Markdowns depend on the marginal revenue product of labor, which is not observable. Therefore, we rely on the estimates of Yeh et al. (2022) for U.S. manufacturing firms. They report an average markdown of 0.35, meaning that workers get paid 65 cents on the marginal dollar. The other three moments, we compute directly from the LBD for the period 2015–2019. Less than 6 percent of firms are high monopsony—defined as having a wage bill share above 10 percent—but they employ 40 percent of private sector workers. The average local labor market—defined as a county-subsector pair—is highly concentrated, with an HHI of 0.237. This value is close to what we would observe with just four equal-sized firms. Panel B shows the moments we target in the historical counterfactual in Section 5.4.

The theoretical model has a unit measure of local labor markets. When we solve the model numerically, we work with $25,000$ markets and truncate the number of firms per market at 100. Given our calibration of the market size distribution, we end up with more than $975,000$ firms in total. Figure C.5 shows the steady-state distribution of the market size and the wage bill share. The market size distribution is bimodal. Duopsonies and effectively competitive markets with a 100 firms are the

[16] These moments are endogenous outcomes of the model that in principle depend on all parameters. Nevertheless, there is a clear pattern of identification.

Table 2: Targeted Moments

| Parameter | Corresponding Moment | Source | Value |
|---|---|---|---|
| *A. Baseline Calibration (2015–2019)* | | | |
| $\eta$ | Average markdown | Yeh et al. (2022) | 0.350 |
| $\zeta$ | High-monopsony employment share | LBD | 0.400 |
| $\xi_z$ | High-monopsony firm share | LBD | 0.058 |
| $\xi_j$ | Mean local HHI | LBD | 0.237 |
| *B. Historical Calibration (1978–1980)* | | | |
| $\xi_z$ | High-monopsony firm share | LBD | 0.083 |
| $\xi_j$ | Mean local HHI | LBD | 0.341 |

two most common. The wage bill distribution is concentrated at the bottom. Most firms have low market power, but there is significant mass throughout the unit interval. Wage bill shares above 0.5 are possible only due to dispersion in firm-level productivity.

### 5.3 Labor Market Power and Monetary Policy

With the calibrated model in hand, we turn to the study of monetary policy.

**Aggregate responses.** Figure 6 shows the impulse responses of macroeconomic aggregates to a 25-basis point (annualized) expansionary monetary policy shock. The price and wage inflation responses are typical of standard New Keynesian models: price inflation rises by 40 basis points (annualized) on impact and decays at a rate similar to the shock. Wage inflation rises about three times as much on impact but is much less persistent, reflecting a front-loaded rise in real wages. Despite the rise in real wages, the aggregate markdown $\mathcal{M}_t = 1 - w_t/(A_t m_t)$ increases substantially, implying that workers get paid less relative to the marginal revenue they generate. The reason is that inflation erodes the real wages paid by non-adjusting firms while the marginal revenue product of labor increases due to a rise in the intermediate good price. In turn, adjusters increase their markdown on purpose, because desired markdowns increase with market share.

Looking at the responses of quantities in the second row, more differences emerge relative to the canonical New Keynesian model. Output and employment exhibit small hump shapes, even though the model features no consumption habits on the household side. Furthermore, total factor productivity falls endogenously, displaying more persistence than any other macroeconomic variable. Aggregate employment rises by about 0.07 percentage point at its peak, but output peaks less than 0.06 percentage point above steady state. In present value terms, the decline in TFP damps the output response by 25 percent. Similarly to Baqaee et al. (2024), the transmission of monetary policy has a supply-side channel.

**Heterogeneous responses.** The analysis of the stylized model in Section 4 suggests that the fall in TFP is due to small, low-market power firms responding less than large, high-market power firms.

Figure 6: Aggregate Responses to an Expansionary Monetary Policy Shock

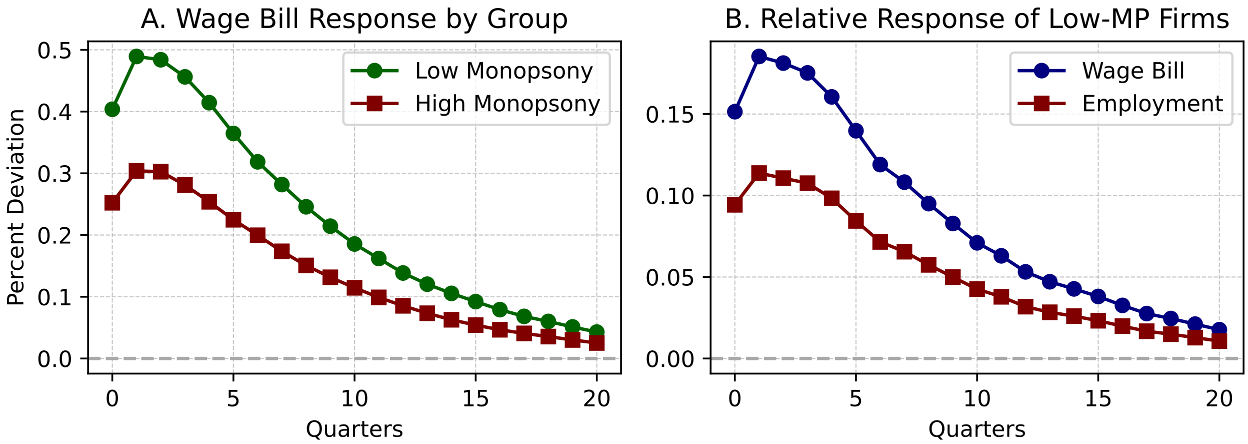

**Note**: Figure shows aggregate responses to an expansionary monetary policy shock from our benchmark New Keynesian model with oligopsonistic wage setting. The first row contains the responses of rates, expressed in percentage point deviations from steady state. The second row contains the responses of quantities, expressed as percentage deviation from steady state.

Next, we demonstrate that the quantitative model well approximates the differential passthrough between low- and high-monopsony firms. Figure 7 panel A shows the real wage bill response of the two groups of firms, using the same definitions as in our empirical analysis. Both low- and high-monopsony firms increase their wage bill on average, but low-monopsony firms respond about 60 percent more than high-monopsony firms in the first year after the shock. Panel B plots the relative response of low-monopsony firms, the same object that we estimate in our main specification and show on Figure 2. As in the data, low-monopsony firms respond more both in terms of their employment and their average wage. In terms of magnitudes, the wage bill of low-monopsony firms responds almost 20 basis points more than that of high-monopsony firms one year after the shock. This is about half of our baseline estimate and close to the specification that controls for the effect of firm age on passthrough. All in all, the model does well in terms of magnitudes. The main discrepancy is in terms of timing. In the data, wage bills rise sluggishly and peak only 2-3 years after the shock. In the model, the response peaks 3 quarters after the shock.[17]

**Partial passthrough and misallocation.** Next, we consider a counterfactual that isolates the effects of oligopsonistic wage setting on monetary transmission. As Section 4 shows, oligopsony works through two channels, both of which operate through endogenous changes in firms' desired markdown.

[17]Enriching the model with information frictions along the lines of Auclert et al. (2020) would make wage responses more sluggish. We decided not to pursue that route because it is orthogonal to the core mechanisms we emphasize.

Figure 7: Impulse Responses of Low- vs High-Monopsony Power Firms




**Note**: Panel A shows the average wage bill responses of low- and high-monopsony power firms to a 25 basis points expansionary monetary policy shock, expressed as percentage deviation from steady state. Panel B shows the wage bill and employment responses of low-monopsony-power firms relative to high-monopsony-power firms. As in the data, both groups of firms expand, but low-monopsony firms gain market share.

Firms that can raise their wages gain market share at the expense of competitors whose wage is stuck. Adjusters' desired markdown increases, and passthrough becomes lower on average. This is the partial passthrough channel. Furthermore, the desired markdown of high-market-power firms falls more than the desired markdown of low-market-power firms. The allocation of employment shifts toward low-market power firms, that are less productive in average, generating a fall in aggregate TFP. This is the misallocation channel. Therefore, we set up the counterfactual economy to be identical to the baseline in every respect, except that the desired markdowns of firms are constant. Formally, the wage setting equation (3.20) becomes

$$W^*_{ijt} = \frac{\eta - (\eta - \zeta)\bar{s}_{ij}}{1 + \eta - (\eta - \zeta)\bar{s}_{ij}} \frac{\mathbb{E}_{\mathcal{S}_{ijt}} \sum_{\tau=0}^{\infty} R_{t,t+\tau} n_{ijt+\tau} M_{t+\tau} z_{ij}}{\mathbb{E}_{\mathcal{S}_{ijt}} \sum_{\tau=0}^{\infty} R_{t,t+\tau} n_{ijt+\tau}}, \tag{5.4}$$

where $\bar{s}_{ij}$ is the steady-state market share of firm $ij$.

The "constant desired markdown" model has the same steady state and admits the same DAG representation as our baseline model, which facilitates a clean analysis of the dynamic effects of monetary policy.[18] In both models, solving for local Nash equilibria conditional on national prices gives rise to a firm block of the form $\{m_t, \pi_t\}_{t=0}^{\infty} \mapsto \{\tilde{w}_t, A_t\}_{t=0}^{\infty}$. Figure C.6 compares the Jacobians of the benchmark and "constant desired markdown" versions of the firm block. As expected, the latter implies more responsive wages (no partial passthrough) and approximately zero TFP responses (no misallocation). Both channels of oligopsony are expected to damp the effect of monetary policy on output. All else equal, partial passthrough implies that the sales price $m_t$ has to rise more to induce firms to raise wages and employment to meet a given surge in aggregate demand, which leads to more

[18]Eliminating labor market power altogether would not only change the steady state but also require that wages be flexible; a profound change to the structure of the model that would complicate the analysis of the monetary transmission.

inflation and prompts the central bank to lean against the boom. All else equal, misallocation implies that meeting a given surge in aggregate demand requires a larger increase in employment, which is only possible by paying higher wages to workers. Again, this requires a rise in the price of the intermediate good and generates inflation.

Table 3: Oligopsony and the Real Effects of Monetary Policy

| | Output | Employment | TFP | Markdown |
|---|---|---|---|---|
| Constant Desired Markdown | 1.00 | 1.00 | 0.00 | 4.76 |
| **Oligopsony** | **0.76** | **0.99** | **-0.23** | **5.36** |
| Partial Passthrough | 0.85 | 0.85 | 0.00 | 5.50 |
| Misallocation | 0.93 | 1.13 | -0.20 | 4.65 |

**Note**: Table reports the present value of the cumulative impulse responses of selected aggregate variables to an expansionary monetary policy shock. The responses are expressed in the same units as in Figure 6: percent deviation from steady state for output, employment, and TFP; percentage point deviation for the markdown. We use the output response in the constant-desired-markdown model as numeraire, to make the results easier to interpret.

Table 3 shows how the real effects of monetary policy—on output, employment, TFP, and aggregate markdown—change as we turn on one or both channels of oligopsonistic wage setting. Specifically, the table reports the cumulative impulse response of each variable, discounted with the steady-state real interest rate. The first row sets the reference point: in the model with fixed desired markdowns, an expansionary monetary policy shock increases output by 1 percent (a normalization), reflecting the same increase in employment and no change in TFP.[19] Even though the desired markdown of every firm is fixed by construction, the aggregate markdown rises by 4.76 percent. The real wages of non-adjusters are eroded by inflation while MRPL (which equals $z_{ij}m_t$) increases for all firms. Moreover, even adjusters do not fully pass through the initial rise in MRPL, given that they have to keep that nominal wage for 5 quarters, on average. In the full oligopsony model, output responds 24 percent less, resulting from a 0.99 percent rise in employment and a 0.23 percent decline in TFP. In addition, the aggregate markdown rises 60 basis points more, reflecting the endogenous rise in desired markdowns among adjusters.

One may be be tempted to conclude that the sizable difference in the efficacy of monetary policy is caused exclusively by the misallocation channel. This conclusion would be wrong, however. To isolate the role of partial passthrough, we solve a version of the model in which the Jacobians $\mathcal{J}^{\tilde{w},m}, \mathcal{J}^{\tilde{w},\pi}$ come from the oligopsony model but $\mathcal{J}^{A,m}, \mathcal{J}^{A,\pi}$ come from the constant-markdown model. The third row labeled "partial passthrough" shows that turning on partial passthrough alone damps the output response by 16 percent compared to the constant-markdown model. Similarly, the last row labeled "misallocation" shows turning on misallocation alone damps the output response by 8 percent. This result is a combination of a stronger expansion of employment and a similar decline in TFP as in the full oligopsony model. Notably, the misallocation channel alone damps the rise of the average markdown, which is what enables real wages and employment to rise by more than in the constant-desired-

[19] Our chosen measures of the employment and TFP responses sum up to the output response because the aggregate production function $Y_t = A_t N_t$ is loglinear.

markdown model. The intuition is that employment is reallocated towards low-market-power firms that are not only less productive but also have lower markdowns, which offsets some of the efficiency losses. In sum, both misallocation and partial passthrough contribute meaningfully to the damped output response under oligopsony. Partial passthrough accounts for $(1-0.85)/(1-0.76) \approx 63\%$ of the effect of shutting down both channels simultaneously. Misallocation accounts for $(1-0.93)/(1-0.76) \approx 30\%$. The remaining 7% is attributable to their interaction, as this procedure does not yield an exact decomposition because the inputs of the firm block, $\{m_t, \pi_t\}$, are endogenous variables.

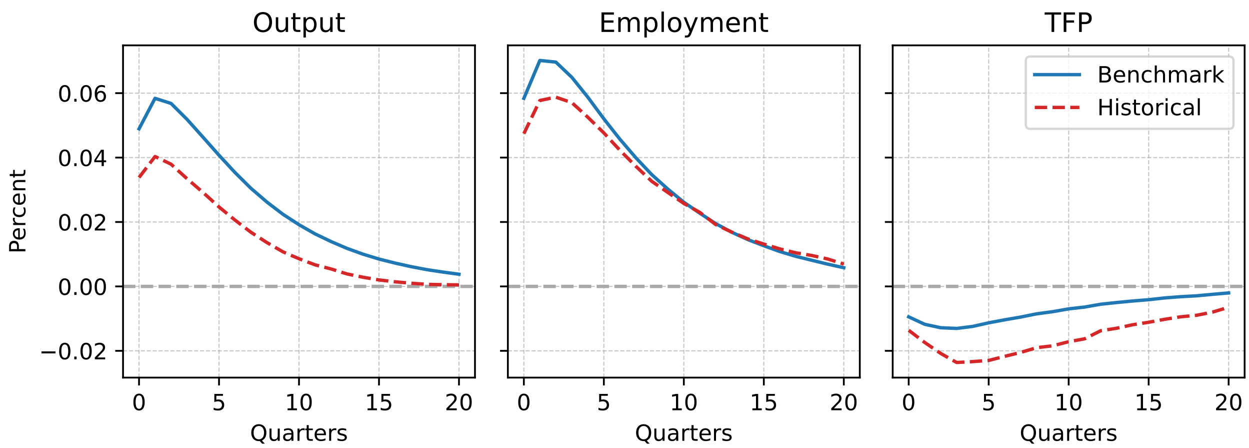
### 5.4 Declining Local Concentration

Local labor market concentration in the United States has declined substantially over the last four decades. Figure C.4 documents this trend using two different measures, both of which are explicitly targeted by our model. The left panel shows that the share of high-monopsony firms has been declining steadily from almost 9 percent in 1978 to less than 6 percent by 2015. The right panel shows that the average HHI across local labor markets had fallen from 0.35 in 1978 to about 0.23 in 2000 and has moved sideways since then. In the last part of the paper, we use our oligopsonistic New Keynesian model to assess the impact of the observed decline in local concentration on the efficacy of monetary policy.

To represent the higher level of labor marker concentration at the beginning of our sample, we recalibrate the shape parameters of the two Pareto distributions governing firm-level productivity and the number of firms per market to match the corresponding moments in the LBD for 1978-1980. The precise values of these targets are shown on panel B of Table 2. We keep the primitive labor supply elasticities $\eta$ and $\zeta$ the same as in the baseline model. Therefore, the two calibrations differ only in the distribution of firms. Intuitively, the historical calibration places more firms on the steeper, more convex part of the desired markdown curve shown on Figure 3. All else equal, this implies larger increases in the desired markdowns of adjusters and reduces passthrough in the aggregate. In addition, the historical distribution features higher productivity dispersion across firms, which amplifies the misallocation channel.

Figure 8 compares the macro responses to an expansionary monetary policy shock under the benchmark and historical calibrations of the model. The output response peaks 40 percent lower and decays faster in the historical calibration than in the benchmark. In present value terms, the output response is less than half of that in the baseline model. The difference is smaller in terms of employment, which rises about 10 percent less in the historical calibration. However, this also means that the differences in TFP responses are even larger, with a cumulative TFP drop that is more than twice as large in the historical calibration than in the benchmark. All in all, shifting the distribution of firms to be consistent with the levels of local concentration seen in 1980 severely reduces the efficacy of monetary policy in stimulating employment and output.

Figure 8: Macro Responses in Historical Counterfactual




# 6 Conclusion

This paper studies how local labor market concentration affects the transmission of monetary policy to macroeconomic aggregates. Using administrative U.S. Census data, we provide evidence that firms respond differently to an identified monetary policy shock according to their wage bill share in the local labor market. Specifically, we estimate that firms with a market share below 10 percent respond about 50 percent more than the firms above that cutoff, both in terms of their wage bill and employment. We propose a New Keynesian model with oligopsonistic competition which can replicate the observed heterogeneity in passthrough. In this framework, heterogeneous passthrough requires (i) nominal wage stickiness and (ii) that the superelasticity of labor supply is decreasing in the wage bill share. If both conditions are met, then monetary easing raises the desired markdown of firms, damping the passthrough of prices to wages. If, in addition, high-market-share firms are more productive on average, then aggregate TFP falls as well. According to our calibrated model, oligopsonistic wage setting reduces the impact of monetary policy on output by 26 percent, with about two-thirds of the difference attributed to partial passthrough and one-third to misallocation. Finally, our model also implies that the observed decline in local labor market concentration in the U.S. since 1980 has amplified the effect of monetary policy on output and contributed to the flattening of the price Phillips curve.

In our model, monetary policy affects firms indirectly by stimulating aggregate demand. In future work, it would be interesting to extend our analysis by allowing monetary policy to affect firms' financing directly. Furthermore, some of the aggregate implications of labor market power—notably weaker real effects of monetary policy and countercyclical TFP—go against the implications of pure product market oligopoly (Wang and Werning, 2022; Baqaee et al., 2024). Studying labor and product market power in a unified framework is another exciting avenue for future research.

## References


Alexander, P. D., L. Han, O. Kryvtsov, and B. Tomlin (2024): "Markups and Inflation in Oligopolistic Markets: Evidence from Wholesale Price Data," Bank of canada staff working paper, 2024-20. 1

Atkeson, A. and A. Burstein (2008): "Pricing-to-Market, Trade Costs, and International Relative Prices," *American Economic Review*, 98, 1998–2031. 3, 3.6, 5.1

Auclert, A., B. Bardóczy, M. Rognlie, and L. Straub (2021): "Using the Sequence-Space Jacobian to Solve and Estimate Heterogeneous-Agent Models," *Econometrica*, 89, 2375–2408. 1, 1, B

Auclert, A., M. Rognlie, and L. Straub (2020): "Micro Jumps, Macro Humps: Monetary Policy and Business Cycles in an Estimated HANK Model," NBER Working Paper 26647. 5.2, 17

Azar, J., I. Marinescu, M. Steinbaum, and B. Taska (2020): "Concentration in US Labor Markets: Evidence from Online Vacancy Data," *Labour Economics*, 66, 101886. 1

Baqaee, D. R., E. Farhi, and K. Sangani (2024): "The Supply-Side Effects of Monetary Policy," *Journal of Political Economy*, 132, 1065–1112. 1, 4.2, 5.3, 6

Barnichon, R. and C. Brownlees (2019): "Impulse Response Estimation by Smooth Local Projections," *Review of Economics and Statistics*, 101, 522–530. 2.2

Bauer, M. D. and E. T. Swanson (2023): "An Alternative Explanation for the Fed Information Effect," *American Economic Review*, 113, 664–700. **??**, 2.3, **??**

Benmelech, E., N. K. Bergman, and H. Kim (2022): "Strong Employers and Weak Employees," *Journal of Human Resources*, 57, S200–S250. 1

Berger, D., K. Herkenhoff, and S. Mongey (2022): "Labor Market Power," *American Economic Review*, 112, 1147–93. 1, 1, 1, 3, 5, 3.1, 3.3, 3.6

Boppart, T., P. Krusell, and K. Mitman (2018): "Exploiting MIT Shocks in Heterogeneous-Agent Economies: The Impulse Response as a Numerical Derivative," *Journal of Economic Dynamics and Control*, 89, 68–92. 1

Burdett, K. and D. T. Mortensen (1998): "Wage Differentials, Employer Size, and Unemployment," *International Economic Review*, 257–273. 1

Caglio, C. R., R. M. Darst, and Kalemli-Özcan (2021): "Collateral Heterogeneity and Monetary Policy Transmission: Evidence from Loans to SMEs and Large Firms," NBER Working Paper 28685. 1

Card, D., A. R. Cardoso, J. Heining, and P. Kline (2018): "Firms and Labor Market Inequality: Evidence and Some Theory," *Journal of Labor Economics*, 36, pp. S13–S70. 1

Champagne, J. and É. Gouin-Bonenfant (2022): "Monetary Policy, Credit Constraints and SME Employment," Bank of Canada Staff Working Paper, 2022-49. 1

Chan, M., S. Salgado, and M. Xu (2023): "Heterogeneous Passthrough from TFP to Wages," *Available at SSRN 3538503*. 1, 2.3

Christiano, L. J., M. Eichenbaum, and C. L. Evans (2005): "Nominal Rigidities and the Dynamic Effects of a Shock to Monetary Policy," *Journal of Political Economy*, 113, 1–45. 5.2

Cloyne, J., C. Ferreira, M. Froemel, and P. Surico (2023): "Monetary Policy, Corporate Finance, and Investment," *Journal of the European Economic Association*, 21, 2586–2634. 1

Durante, E., A. Ferrando, and P. Vermeulen (2022): "Monetary Policy, Investment and Firm Heterogeneity," *European Economic Review*, 148, 104251. 1

Gertler, M. and S. Gilchrist (1994): "Monetary Policy, Business Cycles, and the Behavior of Small Manufacturing Firms," *Quarterly Journal of Economics*, 109, 309–340. 1

Grigsby, J., E. Hurst, and A. Yildirmaz (2021): "Aggregate Nominal Wage Adjustments: New Evidence from Administrative Payroll Data," *American Economic Review*, 111, 428–471. 5.2

Hazell, J. and B. Taska (forthcoming): "Downward Rigidity in the Wage for New Hires," *American Economic Review*. 5.2

Jarociński, M. and P. Karádi (2020): "Deconstructing Monetary Policy Surprises—The Role of Information Shocks," *American Economic Journal: Macroeconomics*, 12, 1–43. 2.1, 2.3

Jordà, Ò. (2005): "Estimation and Inference of Impulse Responses by Local Projections," *American economic review*, 95, 161–182. 1, 2.2

Kaplan, G., B. Moll, and G. L. Violante (2018): "Monetary Policy According to HANK," *American Economic Review*, 108, 697–743. 1

Kimball, M. (1995): "The Quantitative Analytics of the Basic Neomonetarist Model," *Journal of Money, Credit and Banking*, 27, 1241–1277. 3.6, 4.1

Lamadon, T., M. Mogstad, and B. Setzler (2022): "Imperfect Competition, Compensating Differentials, and Rent Sharing in the US Labor Market," *American Economic Review*, 112, 169–212. 1

Manning, A. (2003): *Monopsony in Motion*, Princeton University Press. 1

Meier, M. and T. Reinelt (2024): "Monetary Policy, Markup Dispersion, and Aggregate TFP," *The Review of Economics and Statistics*, 106, 1012–1027. 1

Mongey, S. (2021): "Market Structure and Monetary Non-Neutrality," Working Paper 29233, National Bureau of Economic Research. 1, 3, 3.6, 4.2

Ottonello, P. and T. Winberry (2020): "Financial Heterogeneity and the Investment Channel of Monetary Policy," *Econometrica*, 88, 2473–2502. 1

Reinganum, J. F. and N. L. Stokey (1985): "Oligopoly Extraction of a Common Property Natural Resource: The Importance of the Period of Commitment in Dynamic Games," *International Economic Review*, 26, 161–173. 3.6

Rinz, K. (2022): "Labor market concentration, earnings, and inequality," *Journal of Human Resources*, 57, S251–S283. 1

Robinson, J. (1933): *The Economics of Imperfect Competition*, Palgrave Macmillan. 1

Rossi-Hansberg, E., P.-D. Sarte, and N. Trachter (2021): "Diverging Trends in National and Local Concentration," *NBER Macroeconomics Annual*, 35, 115–150. 1

Ueda, K. (2023): "Duopolistic Competition and Monetary Policy," *Journal of Monetary Economics*, 135, 70–85. 1

Wang, O. and I. Werning (2022): "Dynamic Oligopoly and Price Stickiness," *American Economic Review*, 112, 2815–49. 1, 3, 3.6, 12, 4.2, 6

Yeh, C., C. Macaluso, and B. Hershbein (2022): "Monopsony in the US Labor Market," *American Economic Review*, 112, 2099–2138. 1, 2, 1, 5.2, 2

Yun, T. (2005): "Optimal monetary policy with relative price distortions," *American Economic Review*, 95, 89–109. 13

# Online Appendix

## A Mathematical Derivations and Proofs

### A.1 Household's problem

The representative household solves

$$\max_{\{c_{kt}, n_{ijt}, B_t\}} \mathbb{E}_0 \sum_{t=0}^{\infty} \beta^t U(C_t, N_t) \tag{A.1}$$

$$\text{s.t.} \quad \int_0^1 p_{kt} c_{kt}\, dk + B_t = (1 + i_{t-1}) B_{t-1} + \int_0^1 \sum_{i=1}^{M_j} W_{ijt} n_{ijt}\, dj + \Pi_t, \tag{A.2}$$

$$C_t = \Big( \int_0^1 c_{kt}^{\frac{\epsilon-1}{\epsilon}}\, dk \Big)^{\frac{\epsilon}{\epsilon-1}}, \quad N_t = \Big( \int_0^1 n_{jt}^{\frac{\zeta+1}{\zeta}}\, dj \Big)^{\frac{\zeta}{\zeta+1}}, \quad n_{jt} = \Big( \sum_{i=1}^{M} n_{ijt}^{\frac{\eta+1}{\eta}} \Big)^{\frac{\eta}{\eta+1}}. \tag{A.3}$$

We solve this problem in four steps.

**Step 1.** Derive retailer-specific demand curves and aggregate price index from a static expenditure minimization problem.

$$P_t C_t = \min_{\{c_{kt}\}} \int_0^1 p_{kt} c_{kt} dk \qquad \text{s.t.} \left( \int_0^1 c_{kt}^{\frac{\epsilon-1}{\epsilon}} dj \right)^{\frac{\epsilon}{\epsilon-1}} = \bar{U} \tag{A.4}$$

Let $\lambda_t$ denote the Lagrange multiplier on the constraint. The FOC is

$$p_{kt} = \lambda_t C_t^{\frac{1}{\epsilon}} c_{kt}^{-\frac{1}{\epsilon}} \tag{A.5}$$

Integrating across varieties yields

$$\int_0^1 p_{kt} c_{kt} dk = \lambda_t C_t^{\frac{1}{\epsilon}} \int_0^1 c_{kt}^{\frac{\epsilon-1}{\epsilon}} dk = \lambda_t C_t \tag{A.6}$$

Combining this last last expression with the definition of the price index as $P_t C_t = \int_0^1 p_{kt} c_{kt} dk$ implies that $P_t = \lambda_t$. Substitute this into the FOC to get the retailer-specific demand curve

$$c_{kt} = \left( \frac{p_{kt}}{P_t} \right)^{-\epsilon} C_t \tag{A.7}$$

Substitute the demand curve back into the FOC to express the price index

$$p_{kt}c_{kt} = p_{kt}^{1-\epsilon} P_t^{\epsilon} C_t \tag{A.8}$$

$$\int_0^1 p_{kt}c_{kt}dk = P_t^{\epsilon} C_t \int_0^1 p_{kt}^{1-\epsilon} dk \tag{A.9}$$

$$P_t C_t = P_t^{\epsilon} C_t \int_0^1 p_{kt}^{1-\epsilon} dk \tag{A.10}$$

$$P_t = \left( \int_0^1 p_{kt}^{1-\epsilon} dk \right)^{\frac{1}{1-\epsilon}} \tag{A.11}$$

**Step 2.** Derive local labor-market specific labor supply and aggregate wage index from static income maximization problem.

$$W_t N_t = \max_{\{n_{jt}\}} \int_0^1 W_{jt} n_{jt} dj \qquad \text{s.t.} \left( \int_0^1 n_{jt}^{\frac{\zeta+1}{\zeta}} dj \right)^{\frac{\zeta}{\zeta+1}} = N_t \tag{A.12}$$

Let $\lambda_t$ denote the Lagrange multiplier on the constraint. The FOC is

$$W_{jt} = \lambda_t N_t^{-\frac{1}{\zeta}} n_{jt}^{\frac{1}{\zeta}} \tag{A.13}$$

Integrating across local labor markets yields

$$\int_0^1 W_{jt} n_{jt} dj = \lambda_t N_t^{-\frac{1}{\zeta}} \int_0^1 n_{jt}^{\frac{\zeta+1}{\zeta}} dj = \lambda_t N_t \tag{A.14}$$

Combining this last last expression with the definition of the aggregate wage index as $W_t N_t = \int_0^1 W_{jt} n_{jt} dj$ implies that $W_t = \lambda_t$. Substitute this into the FOC to get the local labor market-specific labor supply curve

$$n_{jt} = \left( \frac{W_{jt}}{W_t} \right)^{\zeta} N_t \tag{A.15}$$

Substitute the labor supply curve back into the FOC to express the aggregate wage index

$$W_{jt} n_{jt} = W_{jt}^{1+\zeta} W_t^{-\zeta} N_t \tag{A.16}$$

$$\int_0^1 W_{jt} n_{jt} dj = W_t^{-\zeta} N_t \int_0^1 W_{jt}^{1+\zeta} dj \tag{A.17}$$

$$W_t N_t = W_t^{-\zeta} N_t \int_0^1 W_{jt}^{1+\zeta} dj \tag{A.18}$$

$$W_t = \left( \int_0^1 W_{jt}^{1+\zeta} dj \right)^{\frac{1}{1+\zeta}} \tag{A.19}$$

**Step 3.** Derive firm-specific labor supply and local wage indexes from static income maximization problem.

$$W_{jt}n_{jt} = \max_{\{n_{ijt}\}} \sum_{i=1}^{M_j} W_{ijt}n_{ijt} \qquad \text{s.t.} \left(\sum_{i=1}^{M} n_{ijt}^{\frac{\eta+1}{\eta}}\right)^{\frac{\eta}{\eta+1}} = n_{jt} \tag{A.20}$$

Analogously to the second step, we obtain

$$n_{ijt} = \left(\frac{W_{ijt}}{W_{jt}}\right)^{\eta} n_{jt}, \qquad W_{jt} = \left(\sum_{i=1}^{M_j} W_{ijt}^{1+\eta}\right)^{\frac{1}{1+\eta}}. \tag{A.21}$$

Combining (A.15) and (A.21) imply the firm-specific labor supply curve in the main text

$$n_{ijt} = \left(\frac{W_{ijt}}{W_{jt}}\right)^{\eta} \left(\frac{W_{jt}}{W_t}\right)^{\zeta} N_t \tag{A.22}$$

**Step 4.** Derive aggregate consumption, labor supply, and savings choices from dynamic utility maximization problem.

$$\max_{C_t,N_t,B_t} \mathbb{E}_0 \sum_{t=0}^{\infty} \beta^t U(C_t, N_t) \qquad \text{s.t. } P_tC_t + B_t = (1 + i_{t-1})B_{t-1} + W_tN_t + \Pi_t \tag{A.23}$$

Let $\lambda_t$ denote the Lagrange multipliers on the budget constraint for all $t \geq 0$. The FOCs are

$$[C_t]: \qquad 0 = U_C(C_t, N_t) - \lambda_t P_t \tag{A.24}$$

$$[N_t]: \qquad 0 = U_N(C_t, N_t) + \lambda_t W_t \tag{A.25}$$

$$[B_t]: \qquad 0 = -\lambda_t + \beta\mathbb{E}_t[(1 + i_t)\lambda_{t+1}] \tag{A.26}$$

Combining the FOCs yields the aggregate Euler equation and labor supply curves

$$U_C(C_t, N_t) = \beta\mathbb{E}_t\left[\frac{1 + i_t}{\pi_{t+1}} U_C(C_{t+1}, N_{t+1})\right], \qquad \frac{W_t}{P_t} = -\frac{U_N(C_t, N_t)}{U_C(C_t, N_t)} \tag{A.27}$$

## A.2 Retailers

Firm $k$ solves

$$\max_{p_{kt}^*,\{c_{kt+\tau},y_{kt+\tau}\}} \sum_{\tau=0}^{\infty} \theta_p^{\tau} R_{t,t+\tau} \left(p_{kt}^* c_{kt+\tau} - M_{t+\tau} y_{kt+\tau}\right) \tag{A.28}$$

$$\text{s.t. } c_{kt+\tau} = y_{kt+\tau} \tag{A.29}$$

$$c_{kt+\tau} = \left(\frac{p_{kt}^*}{P_{t+\tau}}\right)^{-\epsilon} C_{t+\tau} \tag{A.30}$$

Substitute the constraints into the objective function

$$\sum_{\tau=0}^{\infty} R_{t,t+\tau} C_{t+\tau} P_{t+\tau}^{\epsilon} \left[ (p_{kt}^{*})^{1-\epsilon} - M_{t+\tau} (p_{kt}^{*})^{-\epsilon} \right]$$

The FOC wrt $p_t^*$ is

$$0 = \sum_{\tau=0}^{\infty} \theta_p^{\tau} R_{t,t+\tau} C_{t+\tau} P_{t+\tau}^{\epsilon} \left[ (1-\epsilon)(p_{kt}^{*})^{-\epsilon} + \epsilon M_{t+\tau} (p_{kt}^{*})^{-\epsilon-1} \right] \tag{A.31}$$

$$0 = \sum_{\tau=0}^{\infty} \theta_p^{\tau} R_{t,t+\tau} C_{t+\tau} P_{t+\tau}^{\epsilon} \left[ (1-\epsilon) p_{kt}^{*} + \epsilon M_{t+\tau} \right] \tag{A.32}$$

$$p_{kt}^{*} = \frac{\epsilon}{\epsilon - 1} \frac{\sum_{\tau=0}^{\infty} \theta_p^{\tau} R_{t,t+\tau} P_{t+\tau}^{\epsilon} C_{t+\tau} M_{t+\tau}}{\sum_{\tau=0}^{\infty} \theta_p^{\tau} R_{t,t+\tau} P_{t+\tau}^{\epsilon} C_{t+\tau}} \tag{A.33}$$

Given that Calvo shocks are iid and the price index is given by

$$P_t = \left( \int_0^1 p_{kt}^{1-\epsilon} dk \right)^{\frac{1}{1-\epsilon}} \tag{A.34}$$

we can write the law of motion for the price level as

$$P_t = \left[ (1-\theta_p)(p_{kt}^{*})^{1-\epsilon} + \theta_p P_{t-1}^{1-\epsilon} \right]^{\frac{1}{1-\epsilon}} \tag{A.35}$$

**Phillips curve.** Divide (A.33) and (A.35) with $P_t$ to eliminate nominal variables

$$\frac{p_{kt}^{*}}{P_t} = \frac{\epsilon}{\epsilon - 1} \frac{\sum_{\tau=0}^{\infty} \theta_p^{\tau} R_{t,t+\tau} \pi_{t,t+\tau}^{\epsilon+1} C_{t+\tau} m_{t+\tau}}{\sum_{\tau=0}^{\infty} \theta_p^{\tau} R_{t,t+\tau} \pi_{t,t+\tau}^{\epsilon} C_{t+\tau}} \tag{A.36}$$

$$\frac{p_t^{*}}{P_t} = \left( \frac{1 - \theta_p \pi_{t-1,t}^{\epsilon-1}}{1-\theta_p} \right)^{\frac{1}{1-\epsilon}} \tag{A.37}$$

where $m_t \equiv M_t / P_t$ denotes real marginal cost. We can express inflation recursively as

$$\frac{K_t}{F_t} = \left( \frac{1 - \theta_p \pi_{t-1,t}^{\epsilon-1}}{1-\theta_p} \right)^{\frac{1}{1-\epsilon}} \tag{A.38}$$

$$K_t = \epsilon C_t m_t + \theta_p \frac{\pi_{t,t+1}^{\epsilon+1}}{1+i_t} K_{t+1} \tag{A.39}$$

$$F_t = (1-\epsilon) C_t + \theta_p \frac{\pi_{t,t+1}^{\epsilon}}{1+i_t} F_{t+1} \tag{A.40}$$

Loglinearize these equations, assuming $\pi_{ss} = 1$

$$\hat{K}_t - \hat{F}_t = \frac{\theta_p}{1-\theta_p}\hat{\pi}_{t-1,t} \tag{A.41}$$

$$\hat{K}_t = (1-\beta\theta_p)\left(\hat{C}_t + \hat{m}_t\right) + \beta\theta_p\left[(\epsilon+1)\hat{\pi}_{t,t+1} + \hat{K}_{t+1} - \beta di_t\right] \tag{A.42}$$

$$\hat{F}_t = (1-\beta\theta_p)\hat{C}_t + \beta\theta_p\left[\epsilon\hat{\pi}_{t,t+1} + \hat{F}_{t+1} - \beta di_t\right] \tag{A.43}$$

Combine these to get the textbook NKPC

$$\frac{\theta_p}{1-\theta_p}\hat{\pi}_{t-1,t} = (1-\beta\theta_p)\hat{m}_t + \beta\theta_p\left[\hat{\pi}_{t,t+1} + \frac{\theta_p}{1-\theta_p}\hat{\pi}_{t,t+1}\right] \tag{A.44}$$

$$\hat{\pi}_{t-1,t} = \frac{(1-\theta_p)(1-\beta\theta_p)}{\theta_p}\hat{m}_t + \beta\hat{\pi}_{t,t+1} \tag{A.45}$$

## A.3 Producers

We start from the decision problem of a generic producer $ij$ that contemplates its wage strategy for every possible state of the world. Then, we show that the wage setting problem is separable across markets and wage spells, which greatly simplifies the optimization problem. Finally, we derive the wage setting equation.

**Full problem.** The state of the world $s_t$ in period $t$ consists of the monetary policy shock $\epsilon_t$ as well as the Calvo shocks for all firms $\{\{\Theta_{ijt}\}_i\}_j$, where $\Theta_{ijt}$ equals one when firm $i$ in sector $j$ can adjust its wage at time $t$, and zero otherwise. We denote by $s^t$ the history of states, $s^t = \{s_0, \ldots, s_t\}$. The problem of an oligopolistic intermediate good producer is

$$\max_{\{W_{ij}(s^t), n_{ij}(s^t), W_j(s^t)\}} \mathbb{E}_0 \sum_{t=0}^{\infty}\sum_{s^t} R(s^t)\left[M(s^t) z_{ij} n_{ij}(s^t) - n_{ij}(s^t) W_{ij}(s^t)\right], \tag{A.46}$$

$$\text{s.t.} \qquad n_{ij}(s^t) = \left(\frac{W_{ij}(s^t)}{W_j(s^t)}\right)^{\eta}\left(\frac{W_j(s^t)}{W(s^t)}\right)^{\zeta} N(s^t), \tag{A.47}$$

$$W_j(s^t) = \left[W_{ij}(s^t)^{1+\eta} + \sum_{i'\neq i} W_{i'j}(s^t)^{1+\eta}\right]^{\frac{1}{1+\eta}}, \tag{A.48}$$

$$W_{ij}(s^t) = W_{ij}(s^{t-1}) \quad \text{if } \Theta_{ijt} = 0. \tag{A.49}$$

Note that the firm takes as given $\{R(s^t), M(s^t), W(s^t), N(s^t), W_{i'j}(s^t)\}$ when choosing their actions.

**Separability across markets.** Lemma 1 shows that a firm's problem depends only on the history of the aggregate state and the Calvo states on their own market. That is, Calvo realizations affect local outcomes, but local markets interact with each other only via aggregate prices and quantities.

Lemma **1.** *Let $s^t_\epsilon = \{\epsilon_0, \ldots, \epsilon_t\}$ denote the history of the aggregate state, and $s^t_j = \{\{\epsilon_0, \{\Theta_{ij0}\}_i\}, \ldots, \{\epsilon_t, \{\Theta_{ijt}\}_i\}\}$ denote the history of the aggregate state and the Calvo states for the specific market $j$*

*where the firm operates. The firm's problem can be rewritten as*

$$
\begin{aligned}
&\max_{\{W_{ij}(s^t_j), n_{ij}(s^t_j), W_j(s^t_j)\}} \mathbb{E}_0 \sum_{t=0}^{\infty} \sum_{s^t} R(s^t_\epsilon) \left[ M(s^t_\epsilon) z_{ij} n_{ij}(s^t_j) - n_{ij}(s^t_j) W_{ij}(s^t_j) \right], \\
&\textit{s.t.} \qquad n_{ij}(s^t_j) = \left( \frac{W_{ij}(s^t_j)}{W_j(s^t_j)} \right)^{\eta} \left( \frac{W_j(s^t_j)}{W(s^t_\epsilon)} \right)^{\zeta} N(s^t_\epsilon), \\
&\qquad\qquad W_j(s^t_j) = \left[ W_{ij}(s^t_j)^{1+\eta} + \sum_{i' \neq i} W_{i'j}(s^t_j)^{1+\eta} \right]^{\frac{1}{1+\eta}}, \\
&\qquad\qquad W_{ij}(s^t_j) = W_{ij}(s^{t-1}_j) \quad \textit{if } \Theta_{ijt} = 0.
\end{aligned}
\tag{A.50}
$$

Proof. Having a continuum of sectors eliminates Calvo uncertainty in the aggregate, leaving only the aggregate state to matter. For the same reason, the relevant state variables for the firm is only the Calvo realizations in its own sector, other than the aggregate state. ■

**Separability across wage spells.** Next, we argue that a firm that is allowed to adjust its wage need only to consider the states of the world until its next wage adjustment.

Given Definition 1, we can partition the firm's problem into all different wage spells:

$$
\max_{\{W_{ij}(\mathcal{S}_{ijt}), n_{ij}(s^{t+\tau}_j), W_j(s^{t+\tau}_j)\}} \mathbb{E}_0 \sum_{t=0}^{\infty} \sum_{\mathcal{S}_{ijt}} \sum_{\tau=0}^{\infty} \sum_{s^{t+\tau} \in \mathcal{S}_{ijt}} R(s^{t+\tau}_\epsilon) \left[ M(s^{t+\tau}_\epsilon) z_{ij} n_{ij}(s^{t+\tau}_j) - n_{ij}(s^{t+\tau}_j) W_{ij}(\mathcal{S}_{ijt}) \right],
\tag{A.51}
$$

$$
\text{s.t.} \qquad n_{ij}(s^{t+\tau}_j) = \left( \frac{W_{ij}(\mathcal{S}_{ijt})}{W_j(s^{t+\tau}_j)} \right)^{\eta} \left( \frac{W_j(s^{t+\tau}_j)}{W(s^{t+\tau}_\epsilon)} \right)^{\zeta} N(s^{t+\tau}_\epsilon),
\tag{A.52}
$$

$$
W_j(s^{t+\tau}_j) = \left[ W_{ij}(\mathcal{S}_{ijt})^{1+\eta} + \sum_{i' \neq i} W_{i'j}(s^{t+\tau}_j)^{1+\eta} \right]^{\frac{1}{1+\eta}},
\tag{A.53}
$$

where we have used that a firm's own wage cannot vary within a wage spell but its employment might be due to aggregate shocks and changes in the wages of its competitors. We make two observations. First, the firm takes as given all payoff-relevant prices and quantities except their own wage. Second, the firm's own wage is a state variable within wage spells, but has no effect on past or future wage spells.

It follows that we may focus on setting an optimal wage $W^*$ for a single generic wage spell $\mathcal{S}_{ijt}$.

$$
\max_{W^*_{ijt}, \{n_{ij}(s^{t+\tau}), W_j(s^{t+\tau})\}} \mathbb{E}_t \sum_{\tau=0}^{\infty} \sum_{s^{t+\tau} \in \mathcal{S}_{ijt}} R(s^{t+\tau}) \left[ M(s^{t+\tau}) z_{ij} n_{ij}(s^{t+\tau}) - n_{ij}(s^{t+\tau}) W^*_{ijt} \right],
\tag{A.54}
$$

$$
\text{s.t.} \qquad n_{ij}(s^{t+\tau}) = \left( \frac{W^*_{ijt}}{W_j(s^{t+\tau})} \right)^{\eta} \left( \frac{W_j(s^{t+\tau})}{W(s^{t+\tau})} \right)^{\zeta} N(s^{t+\tau}),
\tag{A.55}
$$

$$
W_j(s^{t+\tau}) = \left[ \left( W^*_{ijt} \right)^{1+\eta} + \sum_{i' \neq i} W_{i'j}(s^{t+\tau})^{1+\eta} \right]^{\frac{1}{1+\eta}}.
\tag{A.56}
$$

To emphasize the similarity to the standard Calvo price setting problem of the retailers, we fold the dependence on the histories of states into the expectation operator. Let $\mathrm{P}_t(s^{t+\tau})$ denote the probability of state $s^{t+\tau}$ conditional on $s_t$. The wage spell-specific expectation of a generic variable $X$ is defined as

$$\mathbb{E}_{\mathcal{S}_{ijt}}[X_{t+\tau}] \equiv \mathbb{E}_t\left[\sum_{s^{t+\tau}\in\mathcal{S}_{ijt}} \mathrm{P}_t(s^{t+\tau})X(s^{t+\tau})\right]. \tag{A.57}$$

which allows us to write the wage setting problem more compactly as

$$\max_{\{W^*_{ijt}, n_{ijt+\tau}, W_{jt+\tau}\}} \mathbb{E}_{\mathcal{S}_{ijt}} \sum_{\tau=0}^{\infty} R_{t+\tau}\left[M_{t+\tau}z_{ij}n_{ijt+\tau} - n_{ijt+\tau}W^*_{ijt}\right], \tag{A.58}$$

$$\text{s.t.} \qquad n_{ijt+\tau} = \left(\frac{W^*_{ijt}}{W_{jt+\tau}}\right)^{\eta}\left(\frac{W_{jt+\tau}}{W_{t+\tau}}\right)^{\zeta} N_{t+\tau}, \tag{A.59}$$

$$W_{jt+\tau} = \left[\left(W^*_{ijt}\right)^{1+\eta} + \sum_{i'\neq i} W^{1+\eta}_{i'jt+\tau}\right]^{\frac{1}{1+\eta}}. \tag{A.60}$$

**First order conditions.** Let $\mu_{ijt+\tau}$ denote the Lagrange multiplier on labor supply, and $\nu_{ijt+\tau}$ denote the Lagrange multiplier on the wage index. The FOCs are

$$[W^*_{ijt}]: \qquad \mathbb{E}_{\mathcal{S}_{ijt}} \sum_{\tau=0}^{\infty} R_{t,t+\tau}W^*_{ijt}n_{ijt+\tau} = \mathbb{E}_{\mathcal{S}_{ijt}} \sum_{\tau=0}^{\infty}\left[\eta n_{ijt+\tau}\mu_{ijt+\tau} - W^{-\eta}_{jt+\tau}\left(W^*_{ijt}\right)^{1+\eta}\nu_{ijt+\tau}\right] \tag{A.61}$$

$$[W_{jt+\tau}]: \qquad (\eta-\zeta)n_{ijt+\tau}\mu_{ijt+\tau} = \nu_{ijt+\tau}W_{jt+\tau} \tag{A.62}$$

$$[n_{ijt+\tau}]: \qquad R_{t,t+\tau}\left(M_{t+\tau}z_{ij} - W^*_{ijt}\right) = \mu_{ijt+\tau} \tag{A.63}$$

Combine all three

$$\begin{aligned}
\mathbb{E}_{\mathcal{S}_{ijt}} \sum_{\tau=0}^{\infty} R_{t,t+\tau}W^*_{ijt}n_{ijt+\tau} &= \mathbb{E}_{\mathcal{S}_{ijt}} \sum_{\tau=0}^{\infty}\left[\eta n_{ijt+\tau}\mu_{ijt+\tau} - \left(\frac{W^*_{ijt}}{W_{jt+\tau}}\right)^{1+\eta}\nu_{ijt+\tau}W_{jt+\tau}\right] \\
&= \mathbb{E}_{\mathcal{S}_{ijt}} \sum_{\tau=0}^{\infty} n_{ijt+\tau}\mu_{ijt+\tau}\left[\eta - (\eta-\zeta)\left(\frac{W^*_{ijt}}{W_{jt+\tau}}\right)^{1+\eta}\right] \\
&= \mathbb{E}_{\mathcal{S}_{ijt}} \sum_{\tau=0}^{\infty} R_{t,t+\tau}\left(M_{t+\tau}z_{ij}n_{ijt+\tau} - W^*_{ijt}n_{ijt+\tau}\right)\left[\eta - (\eta-\zeta)\left(\frac{W^*_{ijt}}{W_{jt+\tau}}\right)^{1+\eta}\right]
\end{aligned}$$

which we can rearrange to get

$$W^*_{ijt} = \frac{\mathbb{E}_{\mathcal{S}_{ijt}}\sum_{\tau=0}^{\infty} R_{t,t+\tau}n_{ijt+\tau}\left[\eta - (\eta-\zeta)\left(\frac{W^*_{ijt}}{W_{jt+\tau}}\right)^{1+\eta}\right]M_{t+\tau}z_{ij}}{\mathbb{E}_{\mathcal{S}_{ijt}}\sum_{\tau=0}^{\infty} R_{t,t+\tau}n_{ijt+\tau}\left[1+\eta - (\eta-\zeta)\left(\frac{W^*_{ijt}}{W_{jt+\tau}}\right)^{1+\eta}\right]} \tag{A.64}$$

**Wage bill share.** Next, we prove that $\left(\frac{W_{ijt}}{W_{jt}}\right)^{1+\eta}$ is the wage bill share of firm $ij$ in period $t$.

$$s_{ijt} \equiv \frac{W_{ijt}n_{ijt}}{\sum_{l=1}^{M_j} W_{ljt}n_{ljt}} = \frac{W_{ijt}n_{ijt}}{\sum_{l=1}^{M_j} W_{ljt}n_{ijt}\left(\frac{W_{ijt}}{W_{ljt}}\right)^{-\eta}} = \frac{W_{ij}^{1+\eta}}{\sum_l W_{ljt}^{1+\eta}} = \frac{W_{ijt}^{1+\eta}}{W_{jt}^{1+\eta}} \tag{A.65}$$

The first equality is the definition of the local wage bill share. The second equality uses the ratio of labor supply curves to firms $i$ and $l$. The third equality just collects and cancels terms. The fourth equality uses the expression of the local wage index.

## A.4 Special case

This section contains the proofs of Propositions 1 and 2.

**Proof of Proposition 1.** Proof. Conjecture that there exists a constant factor $A > 0$ such that $W_{i,1}^* = AW_{i,0}^*$ for all $i$. The conjecture implies that the local wage index increases by the same factor, $W_{j,1} = AW_{j,0}$, and relative wages stay unchanged, $\frac{W_{i,0}}{W_{j,0}} = \frac{W_{i,1}}{W_{j,1}}$. Thus, wage bill shares $s_{it}$ and desired markdowns $\mu_{it}$ also remain the same: $s_{i,1} = s_{i,0}$ and $\mu_{i,1} = \mu_{i,0}$. Substituting these into equation (4.1) yields

$$W_{i,0}^* = \frac{z_i M_0}{\mu_{i,0}}, \qquad AW_{i,0}^* = \frac{z_i M_1}{\mu_{i,0}}, \tag{A.66}$$

which validates the conjecture for $A = M_1/M_0$. In sum, flexible wages imply complete and uniform passthrough across firms. ■

**Proof of Proposition 2.** Proof. Let's index the adjusting firms with $s$ and $l$ where $z_s < z_l$. Lemma 2 shows that the more productive firm indeed has a "larger" wage bill share than the less productive firm. Suppose by contradiction that $\Delta \ln W_l \geq \Delta \ln W_s$. We can rewrite equation (4.1) as follows:

$$s_i^{\frac{1}{1+\eta}} = \frac{\eta - (\eta - \zeta)s_i}{1 + \eta - (\eta - \zeta)s_i} z_i \frac{M}{W_j}. \tag{A.67}$$

Divide by period 0 equation

$$\Delta s_i^{\frac{1}{1+\eta}} = \Delta \frac{\eta - (\eta - \zeta)s_i}{1 + \eta - (\eta - \zeta)s_i} \Delta \frac{M}{W_j}. \tag{A.68}$$

We want to show that if $\Delta s_l \geq \Delta s_s$, then $\Delta \frac{\eta - (\eta - \zeta)s_l}{1+\eta-(\eta-\zeta)s_l} < \Delta \frac{\eta-(\eta-\zeta)s_s}{1+\eta-(\eta-\zeta)s_s}$. This would lead to a contradiction. Let $x_0$, $x_1$, $y_0$, $y_1$ be the wage shares in period 0 and 1 of the large firm ($x$) and small firm ($y$). Let $\gamma = \eta - \zeta > 0$. Let's write the longer version:

$$\frac{\eta - \gamma x_1}{\eta - \gamma x_0} \frac{1 + \eta - \gamma x_0}{1 + \eta - \gamma x_1} < \frac{\eta - \gamma y_1}{\eta - \gamma y_0} \frac{1 + \eta - \gamma y_0}{1 + \eta - \gamma y_1}. \tag{A.69}$$

We can rearrange this to

$$1-\frac{\gamma(x_1-x_0)}{(\eta-\gamma x_0)(1+\eta-\gamma x_1)}<1-\frac{\gamma(y_1-y_0)}{(\eta-\gamma y_0)(1+\eta-\gamma y_1)}. \tag{A.70}$$

So the condition holds if and only if

$$\frac{x_1-x_0}{y_1-y_0}\frac{\eta-\gamma y_0}{\eta-\gamma x_0}\frac{1+\eta-\gamma y_1}{1+\eta-\gamma x_1}>1. \tag{A.71}$$

Recall that we've assumed $\frac{x_1}{x_0}\geq\frac{y_1}{y_0}$, which implies $\frac{x_1-x_0}{y_1-y_0}\geq\frac{x_0}{y_0}$. So it suffices to show that

$$\frac{x_0}{y_0}\frac{\eta-\gamma y_0}{\eta-\gamma x_0}\frac{1+\eta-\gamma y_1}{1+\eta-\gamma x_1}>1. \tag{A.72}$$

Rearranging, we obtain that the condition holds if

$$\frac{\eta/y_0-\gamma}{\eta/x_0-\gamma}\frac{1+\eta-\gamma y_1}{1+\eta-\gamma x_1}>1. \tag{A.73}$$

The first fraction is strictly greater than one if and only if $x_0>y_0$, and the second fraction is strictly greater than one if and only if $x_1>y_1$. Lemma 2 shows that indeed these two conditions hold.
In addition, we've implicitly assumed that $(y_1-y_0)>0$. That is, that $\Delta s_l>1$. To see that this is the case, recall that we've assumed $\Delta s_l\geq\Delta s_s$. Suppose by contradiction that $\Delta s_l\leq 1$. Our assertion that $\Delta s_l\geq\Delta s_s$, then implies $\Delta s_s\leq 1$. Since the wage of the non-adjuster is unchanged, $\Delta W_m=1$, $\Delta s_l$ and $\Delta s_h$ are lower than one then imply that $\Delta W_l\leq 1$ and $\Delta W_h\leq 1$. This implies that $\Delta W_j\leq 1$. Now consider equation (A.68). The RHS consists of the multiplication of four terms. $\Delta W_j\leq 1$ implies that the third and fourth terms are both greater than one. $\Delta M$ is strictly greater than one. And if $s_l^1<s_s^0$, then also the first term is greater than one. We then get a contradiction, because the RHS then implies that $\Delta s_l>1$. Thus, we have that $\Delta s_l$ is greater than 1, which concludes our proof as it leads to a contradiction. Therefore, we have that $\Delta\ln W_l<\Delta\ln w_s$. ■

LEMMA **2.** *Consider two adjusting firms in sector $j$ with $z_l>z_s$. Then, we have that $n_l>n_s$.*

Proof. Dividing the labor supply curves yields

$$\frac{n_l}{n_s}=\left(\frac{W_l}{W_s}\right)^{\eta}.$$

So we need to show that $W_l>W_s$, which is equivalent to $s_l>s_s$. Rearranging the optimal wage setting equation (4.1), we obtain

$$s_i^{\frac{1}{1+\eta}}=\frac{\eta-(\eta-\zeta)s_i}{1+\eta-(\eta-\zeta)s_i}z_i\frac{M}{W_j}.$$

So that

$$\left(\frac{s_l}{s_s}\right)^{\frac{1}{1+\eta}}=\frac{1-\frac{1}{1+\eta-(\eta-\zeta)s_l}}{1-\frac{1}{1+\eta-(\eta-\zeta)s_s}}\frac{z_l}{z_s}. \tag{A.74}$$

Suppose by contradiction then $s_s \geq s_l$. Then, we have that

$$\frac{1}{1+\eta-(\eta-\zeta)s_l} \leq \frac{1}{1+\eta-(\eta-\zeta)s_s},$$

so that

$$\frac{1-\frac{1}{1+\eta-(\eta-\zeta)s_l}}{1-\frac{1}{1+\eta-(\eta-\zeta)s_s}} \geq 1.$$

However, because $\frac{z_l}{z_s} > 1$, equation (A.74) implies that $s_l > s_s$ because the RHS of the equation is strictly greater than one. A contradiction. Thus, $s_l > s_s$. ■

# B Solution Method

This appendix describes how to compute the Jacobians of the producer block. The Jacobians are conditional on the realization of Calvo shocks, which break the time symmetries that facilitate the fake news algorithm in Auclert et al. (2021). Nevertheless, the Jacobians can be computed quite efficiently, given that the wage setting equation is a well-behaved rootfinding problem with analytical derivatives.

This appendix assumes that the deterministic steady state of the model is known. The steady state is straightforward to compute, for example by iterating the static best responses from an initial guess that all firms set the same markdown.

## B.1 Best Response Functions

The first step is to implement a function that maps the sequences $\{m_t, \pi_t, w_{jt}\}_{t=0}^T$ into best responses $\{\tilde{w}_{ijt}\}_{t=0}^T$ for all firms $i = 1, \ldots, M_j$. We perform this step separately for all markets $j = 1, \ldots, J$, and compute the implied local wage indices $\{\tilde{w}_{jt}\}_{t=0}^T$ as well. The tilde signifies that the best responses are generally different from the input wage indices. Note that we do not include $\{W_t, N_t, i_t\}_{t=0}^T$ among the inputs of the best response function, even though they appear in the first-order condition, because they have no first-order effect on $\tilde{w}_{ijt}$.

**Optimal reset wage.** Let $H_{ijt}$ denote the planning horizon of firm $i$ on market $j$ resetting its wage at time $t$, i.e. the length of wage spell $\mathcal{S}_{ijt}$. Given the assumption of perfect foresight, the wage setting equation (3.20) can be rewritten in real terms as

$$w^*_{ijt} = \frac{\sum_{\tau=0}^{H_{ijt}} R_{t,t+\tau}\pi_{t,t+\tau}n_{ijt+\tau}\left[\eta - (\eta-\zeta)s_{ijt+\tau}\right]m_{t+\tau}z_{ij}}{\sum_{\tau=0}^{H_{ijt}} R_{t,t+\tau}n_{ijt+\tau}\left[1+\eta-(\eta-\zeta)s_{ijt+\tau}\right]}. \tag{B.1}$$

Substituting the labor supply curve and the expression for market share yields

$$w^*_{ijt} = \frac{\sum_{\tau=0}^{H_{ijt}} R_{t,t+\tau}\pi_{t,t+\tau}\left[\eta - (\eta-\zeta)\left(\frac{w^*_{ijt}}{w_{jt+\tau}}\right)^{1+\eta}\right]m_{t+\tau}z_{ij}\left[\left(\frac{w^*_{ijt}}{w_{jt+\tau}}\right)^{\eta}\left(\frac{w_{jt+\tau}}{w_{t+\tau}}\right)^{\zeta}N_{t+\tau}\right]}{\sum_{\tau=0}^{H_{ijt}} R_{t,t+\tau}\left[1+\eta-(\eta-\zeta)\left(\frac{w^*_{ijt}}{w_{jt+\tau}}\right)^{1+\eta}\right]\left(\frac{w^*_{ijt}}{w_{jt+\tau}}\right)^{\eta}\left(\frac{w_{jt+\tau}}{w_{t+\tau}}\right)^{\zeta}N_{t+\tau}}. \tag{B.2}$$

This is a nonlinear equation that we solve for $w^*_{ijt}$ via Newtons' method. Collecting the coefficients on the unknown, we can write

$$w^*_{ijt} = \frac{z_{ij}\cdot(w^*_{ijt})^{\eta}\cdot A_{ijt} - z_{ij}\cdot(w^*_{ijt})^{1+2\eta}\cdot B_{ijt}}{(w^*_{ijt})^{\eta}\cdot C_{ijt} - (w^*_{ijt})^{1+2\eta}\cdot D_{ijt}}, \tag{B.3}$$

where the coefficients are

$$A_{ijt} = \sum_{\tau=0}^{H_{ijt}} R_{t+\tau}\pi_{t,t+\tau}\eta m_{t+\tau}\left[\left(\frac{1}{w_{jt+\tau}}\right)^{\eta}\left(\frac{w_{jt+\tau}}{W_{t+\tau}}\right)^{\zeta} N_{t+\tau}\right], \tag{B.4}$$

$$B_{ijt} = \sum_{\tau=0}^{H_{ijt}} R_{t+\tau}\pi_{t,t+\tau}(\eta-\zeta) m_{t+\tau}\left(\frac{1}{w_{jt+\tau}}\right)^{1+\eta}\left[\left(\frac{1}{w_{jt+\tau}}\right)^{\eta}\left(\frac{w_{jt+\tau}}{W_{t+\tau}}\right)^{\zeta} N_{t+\tau}\right], \tag{B.5}$$

$$C_{ijt} = \sum_{\tau=0}^{H_{ijt}} R_{t+\tau}(1+\eta)\left(\frac{1}{w_{jt+\tau}}\right)^{\eta}\left(\frac{w_{jt+\tau}}{W_{t+\tau}}\right)^{\zeta} N_{t+\tau}, \tag{B.6}$$

$$D_{ijt} = \sum_{\tau=0}^{H_{ijt}} R_{t+\tau}(\eta-\zeta)\left(\frac{1}{w_{jt+\tau}}\right)^{1+2\eta}\left(\frac{w_{jt+\tau}}{W_{t+\tau}}\right)^{\zeta} N_{t+\tau}. \tag{B.7}$$

We precompute these coefficients and use them to evaluate both the objective function and its derivative analytically as

$$F(w_{ijt}^*) = z_{ij}\left[(w_{ijt}^*)^{\eta} A_{ijt} - (w_{ijt}^*)^{1+2\eta} B_{ijt}\right] - \left[(w_{ijt}^*)^{1+\eta} C_{ijt} - (w_{ijt}^*)^{2+2\eta} D_{ijt}\right], \tag{B.8}$$

$$\begin{aligned} F'(w_{ijt}^*) = {} & z_{ij}\left[\eta (w_{ijt}^*)^{\eta-1} A_{ijt} - [1+2\eta]\,(w_{ijt}^*)^{2\eta} B_{ijt}\right] \\ & - \left[(1+\eta)(w_{ijt}^*)^{\eta} C_{ijt} - (2+2\eta)(w_{ijt}^*)^{1+2\eta} D_{ijt}\right], \end{aligned} \tag{B.9}$$

and solve for $w_{ijt}^*$ for all wage spells by Newton's method.

**Realized wage.** Given a solution for the optimal reset wage $w_{ijt}^*$, the actual real wage of firm $i$ on market $j$ can be computed by forward iteration

$$\tilde{w}_{ijt} = \begin{cases} w_{ijt}^* & \text{if } \Theta_{ijt} = 1, \\ \frac{\tilde{w}_{ijt-1}}{1+\pi_t} & \text{if } \Theta_{ijt} = 0. \end{cases} \tag{B.10}$$

**Aggregation.** Given sequences of firm level wages $\{\tilde{w}_{ijt}\}$, we compute the local wage indices as

$$\tilde{w}_{jt} = \left(\sum_{i=1}^{M_j} \tilde{w}_{ijt}^{1+\eta}\right)^{\frac{1}{1+\eta}}. \tag{B.11}$$

To summarize, we have constructed best-response functions for each market in sequence space. Letting bold letters denote sequences, we may write it as $\tilde{\mathbf{w}}_j = \mathbf{F}_j(\mathbf{w}_j, \boldsymbol{\pi}, \mathbf{m})$.

## B.2 Nash Equilibrium on Local Markets

Nash equilibrium is the fixed point $\tilde{\mathbf{w}}_j = \mathbf{w}_j$. By the implicit function theorem, the Jacobian of the Nash-equilibrium wage with respect to either input $\mathbf{x} \in \{\boldsymbol{\pi}, \mathbf{m}\}$, evaluated at the steady state, is

$$\frac{d\mathbf{w}_j}{d\mathbf{x}} = -\left(\frac{\partial \mathbf{F}_j}{\partial \mathbf{w}_j} - \mathbf{I}\right)^{-1} \frac{\partial \mathbf{F}_j}{\partial \mathbf{x}}. \tag{B.12}$$

We compute these "solved Jacobians" separately for each market, and then obtain the Jacobian of the aggregate wage index as

$$\frac{d\mathbf{w}}{d\mathbf{x}} = w_{ss}^{-\zeta}\left(\frac{1}{J}\sum_{j=1}^{J} w_{j,ss}^{\zeta}\frac{d\mathbf{w}_j}{d\mathbf{x}}\right). \tag{B.13}$$

## B.3 Jacobians of TFP

Let us define TFP $A_t$ such that $Y_t = A_t N_t$. Our goal is to compute the Jacobians of $\{A_t\}_{t=0}^{T}$ with respect to $\{\pi_t, m_t\}_{t=0}^{T}$ under Nash equilibrium on local markets. Aggregate output is

$$Y_t = \frac{1}{J}\sum_{j=1}^{J}\sum_{i=1}^{M_j} z_{ij} n_{ijt} = \frac{1}{J}\sum_{j=1}^{J}\sum_{i=1}^{M_j} z_{ij}\left(\frac{w_{ijt}}{w_{jt}}\right)^{\eta}\left(\frac{w_{jt}}{w_t}\right)^{\zeta} N_t. \tag{B.14}$$

Therefore TFP is

$$A_t = \frac{Y_t}{N_t} = \frac{1}{J}\sum_{j=1}^{J}\sum_{i=1}^{M_j} z_{ij}\left(\frac{w_{ijt}}{w_{jt}}\right)^{\eta}\left(\frac{w_{jt}}{w_t}\right)^{\zeta}. \tag{B.15}$$

As in the case of wage strategies, we want to compute a sufficient statistic $A_{jt}$ separately for each market. The natural candidate is

$$A_{jt} = \sum_{i=1}^{M_j} z_{ij}\left(\frac{w_{ijt}}{w_{jt}}\right)^{\eta}\left(\frac{w_{jt}}{w_t}\right)^{\zeta} \tag{B.16}$$

Conceptually, this is a function in sequence space $\mathbf{A}_j(\mathbf{w}_j, \mathbf{w}, \mathbf{m}, \boldsymbol{\pi})$, and we can use the Jacobians of Nash equilibrium wages to compute

$$\frac{d\mathbf{A}_j}{d\mathbf{x}} = \frac{\partial \mathbf{A}_j}{\partial \mathbf{x}} + \frac{\partial \mathbf{A}_j}{\partial \mathbf{w}_j}\frac{d\mathbf{w}_j}{d\mathbf{x}} + \sum_{j=1}^{J}\frac{\partial \mathbf{A}_j}{\partial \mathbf{w}}\frac{d\mathbf{w}}{d\mathbf{x}} \tag{B.17}$$

It simpler and more efficient however, to work with $B_{jt} = A_{jt} w_t^{\zeta}$ whose Jacobian can be computed market by market as

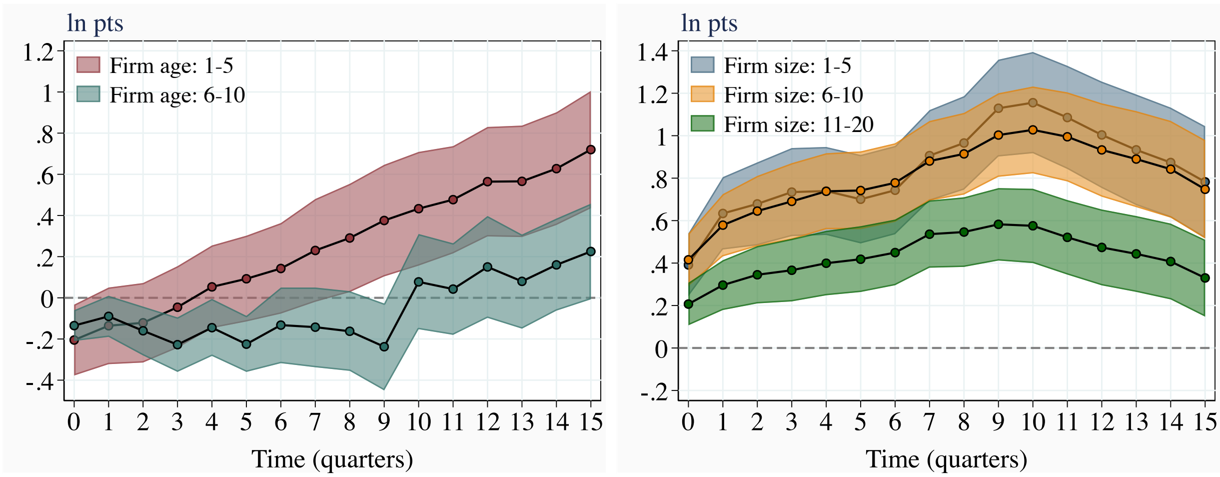
$$\frac{d\mathbf{B}_j}{d\mathbf{x}} = \frac{\partial \mathbf{B}_j}{\partial \mathbf{x}} + \frac{\partial \mathbf{B}_j}{\partial \mathbf{w}_j}\frac{d\mathbf{w}_j}{d\mathbf{x}} \tag{B.18}$$

and then aggregated to obtain

$$\frac{d\mathbf{A}}{d\mathbf{x}} = \frac{1}{J}\sum_{j=1}^{J}\left(w_{ss}^{-\zeta}\cdot\frac{d\mathbf{B}_j}{d\mathbf{x}} - B_{j,ss}\cdot\zeta w_{ss}^{-\zeta-1}\frac{d\mathbf{w}}{d\mathbf{x}}\right). \tag{B.19}$$

# C Additional Figures and Tables

Figure C.1: Heterogenous Response of Firm Wage Bill to a Monetary Policy Shock

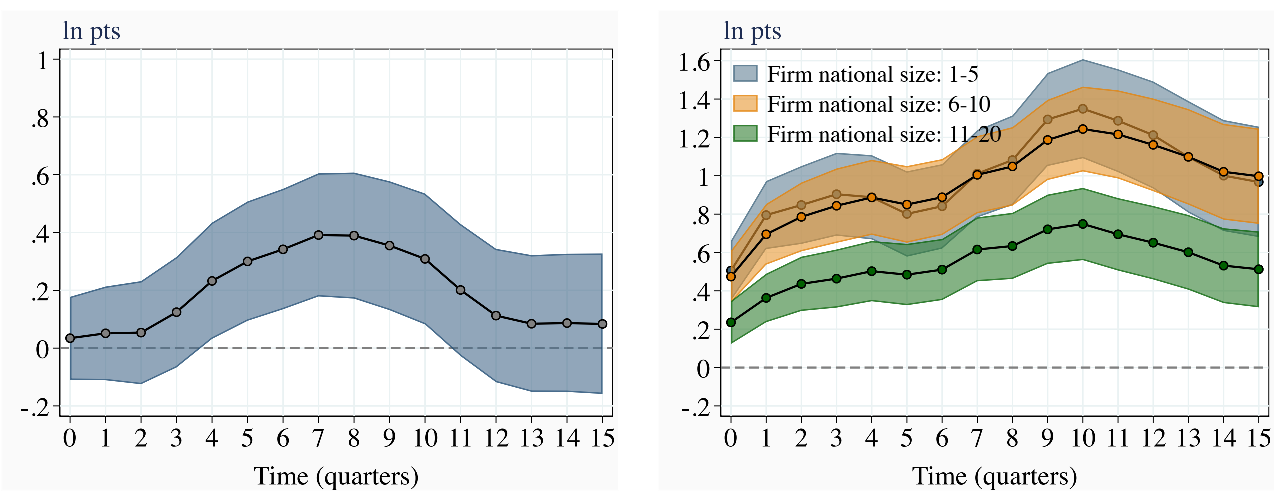



**Note**: Figure shows the relative of the wage bill of firms of different characteristics to a 25 basis points expansionary monetary policy shock. In each regression, we control for the monopsony power of the firm. The left panel displays the response of young and mid age firm relative to older firms (10 years or more at the time of the shock). The right plot shows the response of firms of different (lagged) employment size relative to firms with 20 employees or more in a particular labor market.

Figure C.2: Response of Firm Wage Bill to a Monetary Policy Shock by National Size

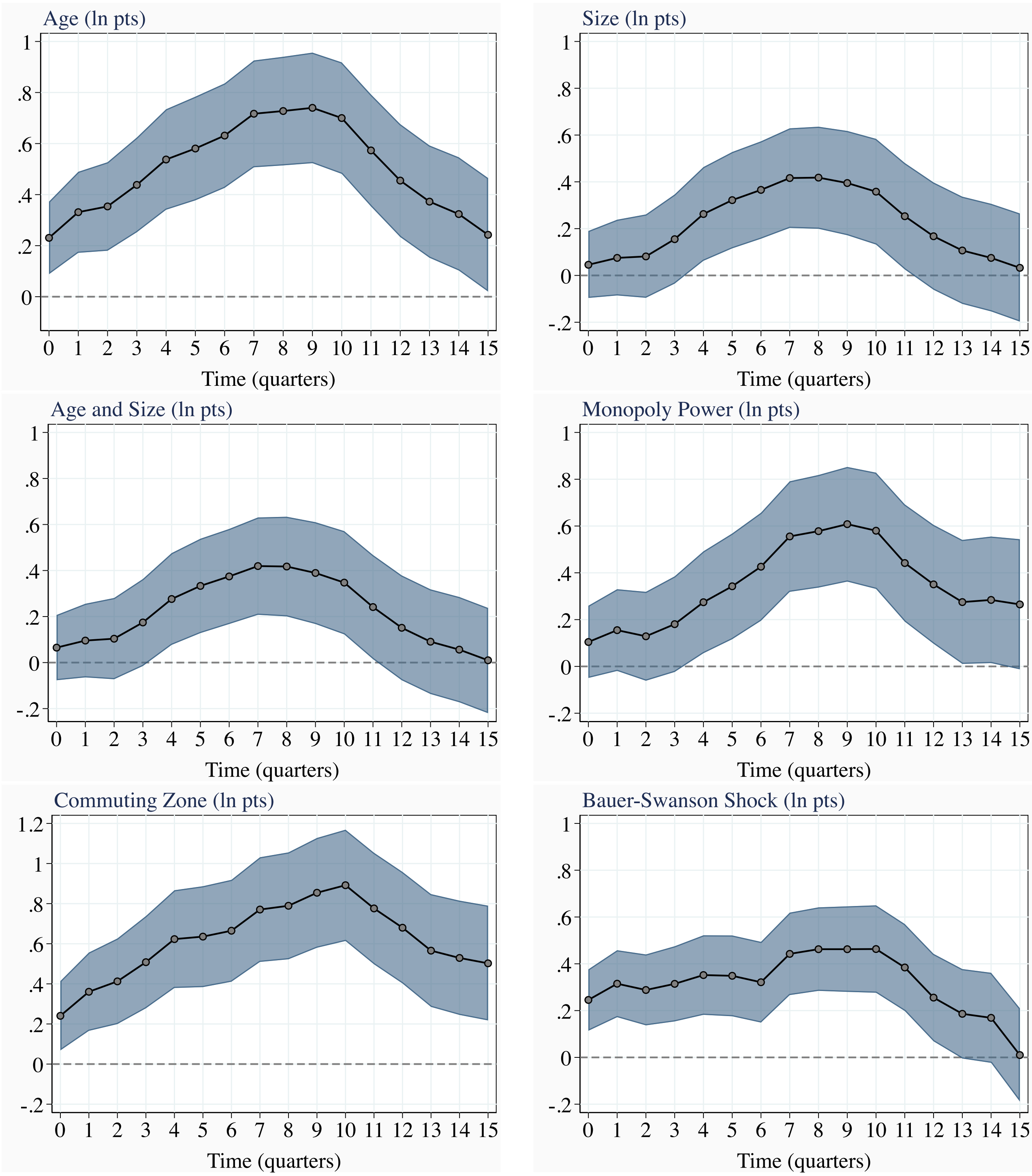
(a) Low monopsony power relative response

(b) Firm size relative response



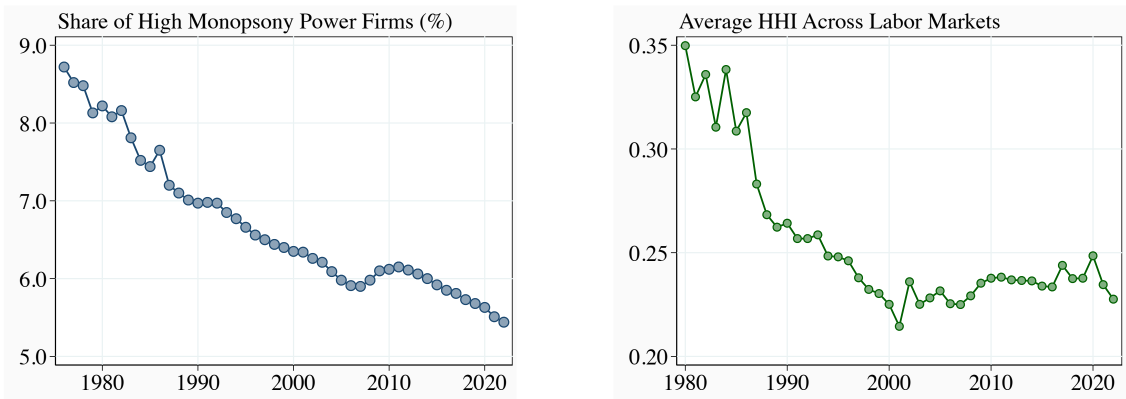

**Note**: Figure shows the response of firms wage bill to a 25 basis points expansionary monetary policy shock. The left panel displays the response of low monopsony power firms relative to high monopsony power firms. The regression includes an interaction term of the monetary shock and a firm national employment size group dummy. The right plot shows the coefficient associated to that interaction for different size groups.

Figure C.3: Full Moving Average Local Projection



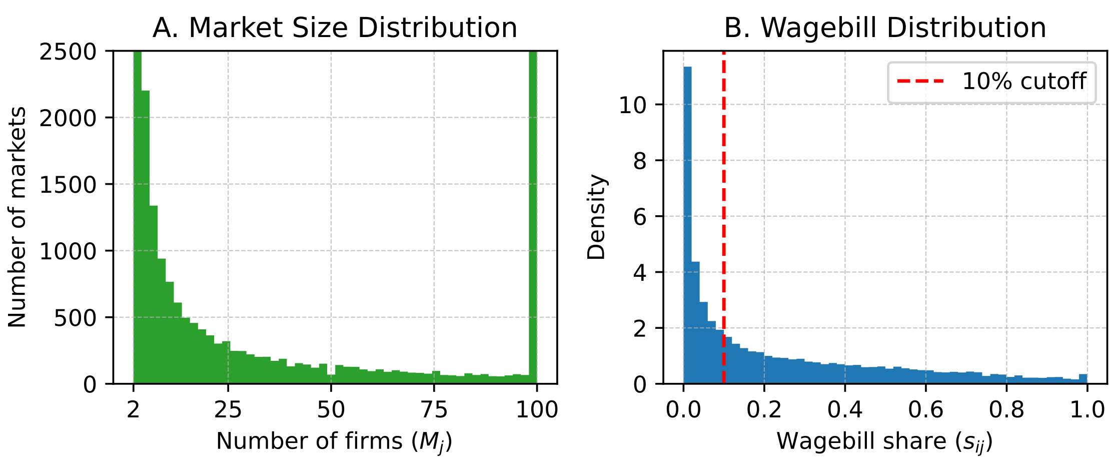

**Note**: Figure shows the response of firms wage bill of low monopsony power firms *relative* to high monopsony power firms to a 25 basis points expansionary monetary policy shock. In each case, we include the interaction of the corresponding control and the monetary policy shock. Shaded areas are 95 percent confidence interval. *Age* indicates firm age controls. *Monopoly* is a dummy equal to 1 if the firm is at the top 5 percent of the within NAICS-4 sales distribution. *Size* indicates firm employment size controls. *CZ labor market* indicates labor market power defined at the commuting zone level. *Bauer-Swanson* indicates Bauer and Swanson (2023) monetary policy shocks.

Figure C.4: Local Labor Market Concentration in the United States



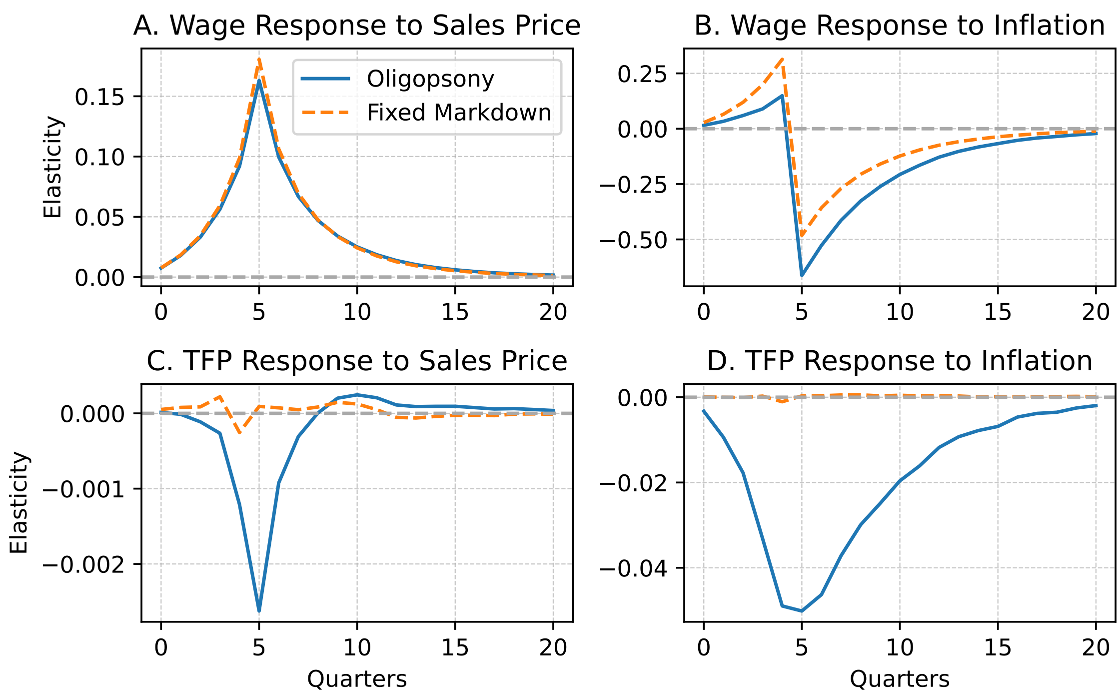

**Note**: Figure shows the evolution of local labor market concentration in the the United States. We report two different measures, both of which are computed from the LBD. The left panel shows the share of high-monopsony firms, defined as having a market share of 10 percent or above. The right panel shows the average HHI across local labor markets.

Figure C.5: Steady-State Distributions




**Note**: Figure shows the distributions of the number of firms per market and the local wagebill shares of firms in the steady state of our model. The model has $10,000$ local markets inhabited by $390,000$ firms. Panel A shows that the number of firms per market range from 2 to a 100. The model as 390,000 firms. Panel B shows that the wagebill share ranges from almost zero to 0.999, with the majority of firms having a wagebill share of less than 10 percent.

Figure C.6: Firm Block Jacobians




**Note**: Figure visualizes the differences between the Jacobians of the firm block $\{m_t, \pi_t\} \mapsto \{\tilde{w}_t, A_t\}$ under oligopsonistic competition and fixed desired markdowns. Each line may be interpreted as an impulse response to a one-time increase in $m_t$ or $\pi_t$ that is announced in period 0. For example, panel A shows $\frac{\partial \tilde{w}_t}{\partial m_5} \frac{m_{ss}}{\tilde{w}_{ss}}$ for $t = 0, 1, \ldots, 20$.