QUANTIFYING MARKET POWER AND BUSINESS DYNAMISM IN THE MACROECONOMY*

Jan De Loecker
KU Leuven†

Jan Eeckhout
UPF Barcelona‡

Simon Mongey
University of Chicago§

March 7, 2022

Abstract

We propose a general equilibrium economy with oligopolistic output markets and endogenous market structure in which two channels can cause a change in market power: (i) technology, via changes in the productivity distribution and the cost of entry, (ii) competition, via changes to the number of potential competitors. First, we show that, when used in conjunction with data on markups and costs, data on business dynamism can be used to disentangle these channels. Second, we estimate that both channels are necessary to explain these data between 1980 and 2016. Third, our estimated changes in technology and market structure over this period yielded positive welfare effects from reallocation and selection, but, quantitatively, these were offset by negative effects from deadweight loss and overhead. Fourth, the changes we identify replicate cross-sectional patterns in declining business dynamism, declining equilibrium wages and labor force participation, and sales reallocation toward larger, more productive firms.

Keywords. Business Dynamism. Market Power in the Aggregate Economy. Technological Change. Market Structure. Reallocation. Endogenous Markups. Wage Stagnation. Labor Share. Passthrough.
JEL. C6. D4. D5. L1.

*We thank Ufuk Akcigit, Philipp Aghion, Steve Berry, Ariel Burstein, Emmanuel Farhi, John Haltiwanger, Tom Holmes, Virgiliu Midrigan, Chad Syverson, John Sutton, John Van Reenen for many useful comments and insightful discussions as well as seminar audiences. De Loecker gratefully acknowledges support from the ERC, Consolidator grant 816638, and Eeckhout from the ERC, Advanced grant 882499, from AEI (Severo Ochoa, Barcelona School of Economics CEX2019-000915-S), and from ECO2015-67655-P. We have benefited from excellent research assistance by Wei Hua, Shubhdeep Deb, Hyejin Park and Renjie Bao.

†jan.deloecker@kuleuven.be
‡jan.eeckhout@upf.edu – ICREA-GSE-CREi
§mongey@uchicago.edu

1 Introduction

There is an ongoing debate about the state of competition in the US economy over the last four decades, and the potential implications of increased market power. A wide range of measures, including price-cost margins and profitability, suggest a change in market power. Over the same period, a number of significant secular trends have been documented. Real wages have stagnated and have not kept track with productivity growth, leading to a much discussed wage-productivity decoupling. Labor market dynamism, in particular measured by the rate of labor reallocation across firms, has decreased, as has the labor share and labor force participation. In this paper, we link these important secular trends in the macroeconomy to the presence of market power in the product market. In particular we use trends in labor market dynamism and the cost structure of firms to disentangle causes behind the increase in market power, and then assess the additional consequences of these changes for welfare, employment and output.

To achieve this we introduce a quantitative framework designed to account for both the *causes* and *consequences* of market power while taking stock of key facts in the micro data. We develop a general equilibrium model of oligopolistic output markets populated with heterogeneous producers facing competitive input markets. Producers are subject to a fixed cost when operating, such that market structure—the number and identity of competing firms—is endogenous. We quantify the model on data from 1980 to 2016, estimating primitives annually. Data on markups and business dynamism jointly identify a change in technology as well as in market structure. These caused a change in market power, which in turn can account for the aforementioned secular trends in the labor market. A priori, the welfare impact of these changes are ambiguous.

To evaluate the welfare impact, our framework distinguishes two structural channels as sources of market power. The first is technological change. In particular, we think of two determinants of technological change: the implicit cost of entry, and the latent productivity distribution across firms in a given market. Technological change can have a positive effect on consumers through reductions in the cost of production, and therefore lower prices. At the same time, some of these gains may be offset because a firm that is considerably more efficient than its competitors can use its dominance to grab market share and extract rents. The significantly higher efficiency of a dominant firm means they can produce/deliver at low cost, but whether these cost savings are passed onto consumers depends on the overall competition in the market. Incomplete passthrough from costs to prices is a central to our model. The second channel is a change in the market structure itself. If there are fewer competitors, under oligopoly, firms set higher markups, leading to the well-known deadweight loss from market power. A change in the effective number of firms can come about from a variety of processes, such as the consolidation of ownership, exit as a consequence of technological change (further magnifying the first channel), or a reduction in entry through entry barriers.

Of course, these processes are invariably dynamic and interact with the technology channel, turning the market structure into an endogenous object as described by the work of Sutton (2001).

Which of these are the root causes of the increase in market power in the US economy are important for assessing its welfare consequences. At face value, and through the lens of our model, the welfare impact is ambiguous. Welfare increases if higher margins reflect increasing market power due to some firms becoming more productive and reallocation of resources to these firms. Welfare decreases if higher margins reflect deadweight losses associated with changes in market structure. And welfare effects are ambiguous if higher margins reflect increasing fixed costs of production: tougher entry requirements improves the selection of firms, but more resources are tied up in overhead. A key insight of the paper is to show that, when filtered through our model, additional data on two other significant trends in the economy over the past four decades can be used in conjunction with increasing markups to discipline these narratives: the declining employment reallocation rate (i.e. reduced *business dynamism*), and changes in the cost-structure of firms. Matching time-series data on these two moments as well as markups, allows us to isolate these different *causes*.

We make three contributions. First, we back out the underlying market structure of the economy by solving a general equilibrium model with micro-level detail, without supposing a particular market definition ex-ante (e.g. firms compete within a NAICS 3- or 4-digit industry). More precisely, given preferences and firm conduct (i.e., entry and pricing decisions), we recover the effective number of potential competitors by matching the model predictions to data moments without using information on the number of competitors or market shares.

This approach combines several insights from the industrial organization literature. First, Bresnahan (1982, 1989) and Berry, Levinsohn, and Pakes (1995) introduce the notion that a model of market-level conduct paired with a demand system delivers markups and marginal costs in equilibrium. Second, Berry and Reiss (2007) and Berry (1992) model entry decisions to recover fixed costs. Third, De Loecker and Warzynski (2012) relies on cost minimization to measure markups using information on technology parameters and data on revenue and variable input expenditures. We blend these three, but depart in a few important ways. Instead of estimating fixed costs, we discipline the ratio of fixed costs in revenue by moments in the data, in a model-consistent fashion. Instead of estimating demand parameters for all product markets in the US economy, we build on Atkeson and Burstein (2008) and assume a macroeconomic model with many markets where heterogenous firms compete strategically in their own, small market. We then calibrate the preference parameters of the macroeconomic model, relying on moments of markups.

The significant advantage of our approach is that we do not have to take a stance on what constitutes a market, how such a definition changes over time, and which firms engage in strategic interaction. This is

beneficial for a number of reasons. First, defining a market is particularly challenging in a macroeconomic setting like ours covering all industries over a long period of time, let alone for a single industry consisting of differentiated products (for a discussion, see Benkard et al. (2021), Syverson (2019) and Berry et al. (2019)). Second, recent work has debated the merits of concentration ratios that are computed using industry classifications that do not line up with actual product markets.[1] It is important to distinguish markets and industries, especially when simple measures such as concentration ratios (such as HHI) are deployed to infer information about market power, either in levels or changes (see Benkard et al., 2021). Third, a separate, and well-known concern relates to the use of HHI measures outside of the homogeneous good Cournot setup (for a discussion see Syverson, 2019; Berry et al., 2019). Finally, in an aggregate economy like ours, changes over time mechanically affect concentration ratios through changes in population and the development of new industries. In response to these issues, our approach, by construction, does not rely on measures that require a market definition such as industry concentration, or market shares. Rather, the market structure and the extent of the market can be thought of as a residual, solved for indirectly by taking the model to the data, and fitting the moments we select.

Our second contribution is to show that in an extensive validation exercise, our micro-founded model of firm behavior accounts for non-targeted secular trends in the macroeconomy. We fit our model to an increase in markups and declining employment reallocation. First, in the aggregate time-series, the model simultaneously quantifies the decline in aggregate wages, labor share, firm entry, and the secular shift in job creation and destruction toward incumbents. Second, in the cross-sectional time-series, the model matches (i) a decomposition of the increase in markups into within- and between- firm components, capturing the reallocation of sales to high productivity, high markup firms; (ii) new facts on the cross-sectional decline in business dynamism: small firms' reallocation rates declined more than large firms. The overarching mechanism that generates the latter is the link between incomplete passthrough of technology shocks into employment and changes market power. Aggregate and cross-sectional employment reallocation rate data suggests that this change in passthrough connects product and labor market outcomes, and that this relationship has changed over time. Incorporating information on employment reallocation into the estimation of the model provides a route to disentangling changes in the economy.

The third contribution comes from identifying and *quantifying* these changes between 1980 and 2016 and their effects on welfare, which are our main results. First, we establish that when combined with time-series on increasing markups and higher fixed cost shares, declining business dynamism can identify changes in primitives reflecting both technology and market structure. We find that fixed costs have

[1]Market definition becomes increasingly complex in the context of global trade, whereby imports and exports need to be accommodated in the context of measuring concentration ratios in US manufacturing (see Amiti and Heise, 2021). Even within a country, additional issues arise for tradeables versus non-tradeables in defining geographic markets (see Rossi-Hansberg et al., 2021).

increased, so has the dispersion of productivity, while competition has weakened. Second, we find that the decline in welfare associated with these changes is 9 percent (with a 10 percent decline in output). The magnitude of this welfare impact is consistent with Baqaee and Farhi (2017). Third, the model allows us to decompose this decline in welfare and output. We find substantial output gains of 5 percent due to technological change (prices are lower, reflecting dominant firms' superior efficiency) but these gains are outweighed by a 15 percent output loss as these firms set higher markups. Due to technological change, firms are substantially more productive, but they do not pass on their efficiency gains to the customer. Fourth, we use the general equilibrium structure of the economy to provide an accounting of how our identified changes in primitive parameters shape five welfare-relevant aggregate *'wedges'*: productivity, selection, overhead, markups, misallocation. We find that the overhead, selection and markup wedges dominate in shaping output, employment and welfare. Fifth, we study how changes in primitives (market structure and technology) determine these wedges. We find that *jointly* understanding changes in market structure and technology is key for understanding mechanisms and the welfare implications. For example, we find that the decline in business dynamism documented by Decker, Haltiwanger, Jarmin, and Miranda (2020) is composed of offsetting forces: changes in competition decrease dynamism which is partially offset by changes in technology that increase dynamism.

Related Literature. Our paper builds on a large literature on market power in the macroeconomy. Our market-level model of endogenous, variable markups builds on Atkeson and Burstein (2008) and is augmented with an entry stage in the style of Berry (1992). The varying markups in our model are driven by the market structure (how many competitors) and the distribution of firm productivities. Because firms compete in small markets, the distribution from which productivities are drawn in conjunction with the number of competitors has aggregate implications. In that sense, our results build on the literature on granular economies that studies the aggregate implications of the distribution of firm productivities: Gabaix (2011), Grassi (2017), Baqaee and Farhi (2017), Acemoglu, Carvalho, Ozdaglar, and Tahbaz-Salehi (2012), Carvalho and Tahbaz-Salehi (2019), Carvalho and Grassi (2015), and Burstein, Carvalho, and Grassi (2019).

One of the main insights of our welfare analysis is that a rise in markups driven by technological change creates a tradeoff between efficiency and deadweight loss. Customers benefit from firms that are more productive as they sell at lower prices. We see that there is a large welfare enhancing effect from reallocation of market share towards those high productivity firms. This is consistent with the superstar effect of Autor et al. (2017). However, the dispersion in productivities also generates a dominant position of those high productivity firms that allows them to exert market power and extract rents from the customer. The heterogeneity in markups is key to this tradeoff between efficiency from reallocation and deadweight loss from market power and leads to large welfare effects that are negative on balance.

Our paper is most closely related to Edmond, Midrigan, and Xu (2019), Baqaee and Farhi (2017) and Akcigit and Ates (2021). Relative to Edmond, Midrigan, and Xu (2019), we seek to identify changes in the economy over time using the structure of our model and data on employment reallocation, markups and fixed costs. Instead, Edmond, Midrigan, and Xu (2019) carefully characterizes, quantifies and decomposes counterfactual transitions between an imperfectly competitive economy and an efficient benchmark.

A similar exercise is the key feature of Baqaee and Farhi (2017), but in a model with sectoral network linkages and exogenous markups. Baqaee and Farhi (2017) construct aggregate *'wedges'* (e.g. markups and misallocation) by aggregating microdata and then compute welfare gains from setting each wedge to its efficient level. Their reduced form approach treats the joint distribution of markups and sales shares as a primitive. Our exercise is different: we take a model with endogenous markups, fix preferences parameters, and then estimate time-series of primitive parameters required to match data on markups, business dynamism and costs. Our structural approach treats parameters as primitives, and the joint distribution of markups and sales shares as endogenous.

Finally, Akcigit and Ates (2021) develop a theory that links a number of stylized facts in the labor market to those on market power. They propose a Schumpeterian growth model and derive a number of model properties that are consistent with those stylized facts.[2] Our approach differs in that instead of an endogenous growth model, we propose a framework with many oligopolistic markets in a large economy, with firm entry, and rich heterogeneity. But most importantly, our main focus is the quantitative exercise that identifies the parameters of technological change and market structure. This permits us to measure the welfare impact, as well as decompose its origins.[3]

We cannot, unfortunately, solve a fully dynamic model in the context of general equilibrium with oligopoly and heterogeneous firms.[4] This precludes the study of the role of endogenous upfront investment that leads to higher productivity, as proposed by Sutton (1991, 2001). This same logic is also confirmed in the case studies of the roll out of the distribution network of large companies such as Walmart (Holmes, 2011) and Amazon (Houde et al., 2017). Nonetheless, by allowing for fixed costs and an endogenous entry, we capture the notion of upfront investment in order to affect competition and market power. We believe that such investments and endogenous entry play a key role in the current information age, where investment is often in intangibles (see also De Ridder, 2019; Crouzet and Eberly, 2019).

[2] In a model, similar in this Schumpeterian spirit, Aghion et al. (2019) show that market power leads to a decline in innovation. Olmstead-Rumsey (2019) provides micro evidence on patent data and shows that market concentration can account for the productivity slowdown.

[3] One important aspect that is absent from our analysis is the role of market power in the input market, i.e., monopsony or oligopsony. A recent burgeoning empirical literature (Azar, Berry, and Marinescu, 2019; Azar, Marinescu, and Steinbaum, 2017; Hershbein, Macaluso, and Yeh, 2020) highlights the effect of monopsony power on wages, both the level and the distribution. Berger, Herkenhoff, and Mongey (2022) estimates a structural model of oligopsony which is used to decompose welfare.

[4] Even in the context of partial equilibrium models, this already proves challenging. For a fully dynamic game with variable markups and firm-level productivity shocks, with only two firms, see Mongey (2017).

Our approach is closely related to the literature on firm dynamics, pioneered by Jovanovic (1982) and Hopenhayn and Rogerson (1993). In the standard firm-dynamics model, business dynamism on the intensive margin of employment reallocation could decline due to increases in adjustment costs. Higher adjustment costs, however, would lead to lower profits. We offer a complementary explanation, sketched out contemporaneously by Decker et al. (2020) and De Loecker et al. (2020). Dynamism declines because passthrough declines in unison with higher markups. Increasing profits over time therefore provide support for the passthrough mechanism we emphasize.

2 Model

We set up and solve a parsimonious model of imperfect competition in a large economy. For expositional purposes, we present a model in which labor is the only input in production.[5] We consider the steady-state of the economy for a fixed set of parameters. In our estimation exercise, we consider a sequence of steady-states, each corresponding to a different year, and indexed by (potentially) different parameters.

2.1 Setup

Environment. Time is discrete.[6] There are two classes of agents: households and firms. Households are identical, consume goods, supply labor, and trade shares in a representative portfolio of all firms in the economy which pays dividends. The measure of households is normalized to one. Firms are organized in a continuum of markets indexed $j \in [0, 1]$. Each market contains M *potential entrant* firms. Of all potential entrants, $M_j \leq M$ firms choose to enter the market and produce. The entering firms are indexed $i \in \{1, \ldots, M_j\}$.[7] Goods are differentiated along two dimensions, first across markets j, then within markets i. A single firm produces a single good indexed ij.

Households. As in Atkeson and Burstein (2008) the utility of consumption of the differentiated final goods is the double Constant Elasticity of Substitution (CES) aggregator of consumption utility from goods within markets and across the continuum of markets. The cross-market elasticity of demand is denoted $\theta > 1$. The within-sector elasticity of demand is denoted $\eta > 1$. These elasticities are ranked $\eta > \theta$ indicating that the household is more willing to substitute goods within a market (Pepsi vs. Coke vs. Dr. Pepper) than across sectors (Soft drinks vs. Laundry detergent vs. Cars).

[5] In Section 3.1, we expand the model to include capital and intermediates and provide a mapping of the extended model to the data.

[6] For notational simplicity we suppress the time subscript t whenever there is no ambiguity.

[7] The number of potential entrants M is the same in all sectors j. We have also considered a distribution of potential entrants $M(j)$ that is market specific. This is in the spirit of Edmond et al. (2015) and Berger et al. (2022) who consider a stochastic number of firms in each sector. The former assumes that the number of firms follows a geometric distribution, while the latter draw from a combination of Pareto distributions. In those papers there is no endogenous entry and hence no potential entrants.

Households discount the future at rate β, have time-separable utility, and derive period utility from consumption adjusted for the disutility of work. The household chooses sequences of consumption of each good c_{ijt}, and labor supply N_t to maximize:

$$\sum_{t=0}^{\infty} \beta^t U\left(C_t - \overline{\varphi}^{-\frac{1}{\varphi}} \frac{N_t^{1+\frac{1}{\varphi}}}{1+\frac{1}{\varphi}}\right) \quad \text{where} \quad C_t = \left[\int_0^1 c_{jt}^{\frac{\theta-1}{\theta}} dj\right]^{\frac{\theta}{\theta-1}} \quad \text{and} \quad c_{jt} = \left[\sum_{i=1}^{M_{jt}} M_{jt}^{-\frac{1}{\eta}} c_{ijt}^{\frac{\eta-1}{\eta}}\right]^{\frac{\eta}{\eta-1}},$$

These preferences $U(C, N)$ are GHH, which removes wealth effects on labor supply as a force in our comparative statics exercise.[8] Utility from consumption of market j goods is normalized by the size of the market M_{jt} in order to neutralize love of variety effects.[9] The household's expenditures are on goods, which are priced p_{ijt}, shares X_t in the portfolio of firms at price Q_t. Income is from labor supply at wage W_t and returns on shares due to their sale and dividends Π_t, giving the budget constraint:[10]

$$\int_0^1 \sum_{i=1}^{M_{jt}} p_{ijt} c_{ijt}\, dj + Q_t X_{t+1} \leq W_t N_t + \Big(Q_t + \Pi_t\Big) X_t.$$

Firms. Firms are heterogeneous in productivity, with Total Factor Productivity (TFP) denoted by z_{ijt}. The production technology is linear in labor where the quantity of output $y_{ijt} = z_{ijt} n_{ijt}$.[11] In this Section there is no capital as an input in production, but firms must employ ϕ units of *fixed* or *overhead* labor, also at wage W_t, in order to produce. Firms face no adjustments costs over time, as we discussed in the introduction. Such adjustment costs would imply state-variables, and doing so with more than two firms is currently computationally intractable.

The productivities z_{ijt} of the M potential firms in market j evolve stochastically. We assume each firm's productivity follows an AR(1) process

$$\log z_{ijt+1} = \rho \log z_{ijt} + \epsilon_{ijt} \quad , \quad \epsilon_{ijt} \sim \mathcal{N}\left(-\frac{1}{2}\frac{\sigma^2}{1+\rho}, \sigma^2\right). \tag{1}$$

The adjustment to the mean of ϵ_{ijt} ensures that the cross-sectional mean of z_{ijt} (in levels) across the entire

[8] In our main quantitative exercises, reduced output—which the household experiences as lower dividends, wages and capital income—would lead to wealth effects on labor supply. We discuss the implications of this in Section 5. It would be straightforward to add this to the model and quantify its effects, especially given the block recursive structure of the model which we describe below.

[9] For the purpose of our analysis, love of variety effects do not impact the distribution of allocations and market shares in equilibrium. We have checked that we estimate the exact same time-series of parameters whether love of variety effects are present or not. Love of variety effects do change the level of output and consumption. However their proper analysis requires taking a stand on (i) how the measure of industries has expanded over time, here we fix this to 1, i.e. $j \in [0, J_t]$, with $J_t = 1$ for all t, (ii) how varieties are valued. For now we view it as simpler to abstract from these effects.

[10] In the full model, the household accumulates and owns the capital stock. The capital stock is rented out to firms at rate R_t and depreciates each period at rate δ. This yields an Euler equation that pins down the equilibrium rental rate of capital: $R_t = 1/\beta - (1-\delta)$.

[11] We have also analyzed the case with decreasing returns to scale $y_{ijt} = z_{ijt} n_{ijt}^{\kappa}$, $0 < \kappa < 1$ which yields comparable results. The recent work by Chiavari (2022) analyzes how returns to scale have changed over time. For now we focus on returns to scale through changes in the importance of variable and fixed factors in production.

economy is invariant to σ and always equal to one.[12] Productivity shocks generate entry and exit from period to period as well as employment reallocation.

Because there are no adjustment cost, the state vector of market j is $\boldsymbol{z}_{jt} = (z_{1jt}, \ldots, z_{Mjt})$. Granularity within each sector causes "market-specific" shocks to emerge from the idiosyncratic shocks faced by firms (Carvalho and Grassi, 2015). However with a continuum of sectors, these sectoral shocks wash out in the aggregate: i.e. the distribution across markets $\boldsymbol{z}_j \sim G(\boldsymbol{z})$ is constant.[13] We further assume no aggregate shocks. These features mean we can drop time subscripts and references to an aggregate state.

Even though firm production decisions are static, the AR(1) process for productivity means we can analyze firm dynamics. Productivity shocks lead employment of incumbents to fluctuate and some firms that were in the market in one period to exit in the next, and vice versa. This allows us to measure employment reallocation via job creation and job destruction, which is a key input into our estimation exercise. In the quantitative exercise, we measure business dynamism between two adjacent periods with identical parameters but different realizations of the shocks. We then investigate how such measures change over time from steady-state to steady-state as parameters vary.

Timing. At the beginning of each period, productivity shocks for the M potential firms in each sector are realized, which determines $\boldsymbol{z}_{jt}$. Given the realization of the shocks, potential firms decide to operate in the market and produce which requires hiring the overhead fixed labor ϕ or stay out and pay zero. Firms that operate in the market then make their production choices. Whether firms operate or not in period t, their productivity evolves to period $t+1$ according to (1).

Market Competition and Equilibrium. The equilibrium concept within each market is a Nash equilibrium. Dropping time subscripts, we first describe the production stage once M_j firms have entered the market. With a finite number of firms in each market, firms exert market power. We model firms' conduct by means of Cournot quantity competition.[14] Indirectly, firms compete with all firms in the economy, including firms in other markets $-j$, but with a continuum of other markets there are no strategic interactions between firms in market j and j'. Each firm is therefore infinitesimally small relative to all firms in other markets and take the price indices of all other markets p_{-j} as fixed.

Within a market j, there is strategic interaction. Firm i chooses its quantity y_{ij}, taking as given the quantities $\mathbf{y}_{-ij}$ of its $M_j - 1$ competitors. Then, given market demand, the firm produces up to its demand curve, delivering the profit function

$$\max_{y_{ij}} \ \pi(y_{ij}, \mathbf{y}_{-ij}) \ = \ \max_{y_{ij}} \ p(y_{ij}, \mathbf{y}_{-ij}, P, Y) y_{ij} - \left(\frac{W}{z_{ij}}\right) y_{ij} - W\phi, \tag{2}$$

[12] Since the mean of z (in logs) is $-(1/2)\sigma^2/(1-\rho^2)$.

[13] By way of comparison, note that in the 'micro-to-macro' shocks literature, idiosyncratic shocks do not wash out in the aggregate. See for example Gabaix (2011), Gaubert and Itskhoki (2016), and Carvalho and Grassi (2015).

[14] Our results also hold under Bertrand competition, with imperfectly substitutable goods. See Appendix A.3

subject to the inverse demand curve $p(y_{ij}, \mathbf{y}_{-ij}, P, Y)$ from household optimality and constant marginal cost W/z_{ij} due to the linear production technology: $y_{ij} = z_{ij}n_{ij}$. Firms solve for the Cournot-Nash equilibrium in their market. This delivers variable labor demand from each firm $n_{ij}\left(\boldsymbol{z}_j, W, P, Y\right)$.

The general equilibrium solution in addition requires feasibility and market clearing. Aggregating firm level labor demand due to variable and fixed inputs delivers the aggregate labor demand curve:

$$N^d\left(W, P, Y\right) = \int_{\mathbb{R}_+^M} \sum_{i=1}^{M_j} n_{ij}\left(\boldsymbol{z}_j, W, P, Y\right)\, dG\left(\boldsymbol{z}_j\right) + \int_0^1 M_j \phi\, dj. \tag{3}$$

During the entry stage, prior to production, the number of competitors M_j is determined in equilibrium. Firms observe their own productivity, as well as those of all competitors $\mathbf{z}_j$, and have rational expectations with respect to (W, P, Y). It is well known that due to the strategic interaction in the production stage, multiple equilibria with entry of different firms may arise when entry is simultaneous. We use an equilibrium selection device following Berry (1992), which we discuss below.

2.2 Solution

Household solution. The solution to the household problem consists of demand functions for each firm's output, and a labor supply condition. Demand for the goods of firm ij is given by[15]

$$c(p_{ij}, \mathbf{p}_{-ij}, P, C) = \left(\frac{p_{ij}}{p_j(p_{ij}, \mathbf{p}_{-ij})}\right)^{-\eta}\left(\frac{p_j(p_{ij}, \mathbf{p}_{-ij})}{P}\right)^{-\theta} C, \tag{4}$$

$$\text{where} \quad p_j(p_{ij}, \mathbf{p}_{-ij}) = \left[\left(\frac{1}{M_j}\right)\sum_{i=1}^{M_j} p_{ij}^{1-\eta}\right]^{\frac{1}{1-\eta}}, \quad P = \left[\int_0^1 {p_j}^{1-\theta} dj\right]^{\frac{1}{1-\theta}}.$$

Aggregate real consumption is C. The allocation of C to market j depends on the level of the sectoral price p_j relative to the aggregate price P. The allocation of expenditure to firm i is then determined by the level of firm i's price p_{ij} relative to p_j. The aggregate price index is the number P such that PC is equal to aggregate expenditure: $PC = \int_0^1 \sum_{i=1}^{M_j} p_{ij}c_{ij}\, dj$.

The household's labor supply curve is obtained by taking first order conditions of (1) with respect to N, and C: $N^S(W) = \overline{\varphi}W^{\varphi}$. GHH preferences imply that labor supply depends only on W. The labor supply curve with constant Frisch elasticity can be interpreted as representing either the intensive margin labor supply of the continuum of workers in the household (hours worked for each worker) or can similarly be derived as the extensive labor supply among the workers (number of members working fixed hours).[16] When we compare the implications of our model to the data we will take the extensive margin labor supply interpretation and look at aggregate employment.

[15] The complete derivation is standard, and can be found in Appendix A.

[16] Formally, if each member of the household supplied either 1 or zero units of labor, and drew a Generalized Extreme Value (GEV) distributed utility cost of working with tail parameter $1/\varphi$, the aggregate labor supply curve would be identical to that which we derive here. See Berger et al. (2022).

Firm solution. We solve the firm entry and production decisions backwards. Given a continuum of markets, firms take as given their beliefs about aggregate outcomes (W, P, Y). In the last stage, for a given set of entrants M_j, firms choose their output y_{ij} taking into account the strategic interaction with the other firms $\mathbf{y}_{-ij}$ in market j. The first order condition of the profit maximization problem (2) are standard:

$$p_{ij}(y_{ij})\left[1+\frac{p'_{ij}(y_{ij})y_{ij}}{p_{ij}(y_{ij})}\right]=W\frac{\partial n_{ij}}{\partial y_{ij}} \quad \rightarrow \quad p_{ij}(y_{ij})\left[1+\frac{1}{\varepsilon_{ij}}\right]=\frac{W}{z_{ij}} \tag{5}$$

where ε_{ij} is the residual demand elasticity, and $\partial y_{ij}/\partial n_{ij}=z_{ij}$ follows from the linear production technology. Equation (5) automatically leads to the Lerner formula: $\dfrac{p_{ij}-mc_{ij}}{p_{ij}}=-\dfrac{1}{\varepsilon_{ij}}$ where $mc_{ij}=\dfrac{W}{z_{ij}}$.

Market equilibrium. In Appendix A we derive the standard result that under CES preferences, the unique Cournot-Nash equilibrium allocation among M_j entrants satisfies:

$$p_{ij} = \mu_{ij}\left(\frac{W}{z_{ij}}\right) \quad , \quad \mu_{ij}=\frac{\varepsilon_{ij}}{\varepsilon_{ij}+1} \tag{6}$$

$$\text{where} \quad \varepsilon_{ij} := -\frac{\partial \log y_{ij}}{\partial \log p_{ij}}\bigg|_{y^*_{-ij}}=\left[s_{ij}\frac{1}{\theta}+\left(1-s_{ij}\right)\frac{1}{\eta}\right]^{-1} \quad , \quad s_{ij}:=\frac{p_{ij}y_{ij}}{\sum_{i=1}^{M_j}p_{ij}y_{ij}}=\frac{1}{M_j}\left(\frac{p_{ij}}{p_j}\right)^{1-\eta}. \tag{7}$$

The optimality conditions are represented via the markup μ_{ij}, defined as price over marginal cost, and s_{ij} is firm i's share of market j revenue. The firm faces a residual demand curve that is iso-elastic with elasticity ε_{ij}, which is itself determined by the revenue share of the firm s_{ij}. [17]

The residual demand elasticity in equation (7) determines the markup and is directly related to the firm's market share. Firms with a higher market share s_{ij} have steeper residual demand and set higher markups. Those firms with a market share close to one have a residual demand elasticity equal to θ, while firms with a market share close to zero have a residual demand elasticity approximately equal to η. The former effectively behave like monopolists within the market, and only take into account the substitution of goods *outside* of market j. Those goods are not close substitutes since $\theta<\eta$. Instead firms that have a small revenue share, face fierce competition from firms *within* market j. Their residual demand curve is flat as the goods within the market are close substitutes, and markups are consequently lower.

The only determinant of market shares and markups is the market j vector of entrant firm productivities: $\boldsymbol{z}_j^*=(z_1,\ldots,z_{Mj})$. Firms that are more productive (high z_{ij}) can sell at lower prices and therefore take a higher market share. Yet, relative to cost, their prices are high, that is, their markups are high: due to their productive efficiency advantage, they also exert more market power and have a high market share. Using the above equations we can explicitly express a firm's market share as a function of its productivity

[17] The Bertrand-Nash equilibrium is identical in the entire system of equations except for the residual demand elasticity. Under Bertrand instead, we have $\varepsilon_{ij}=s_{ij}\theta+\left(1-s_{ij}\right)\eta$.

and the market shares of its competitors as follows:

$$s_{ij} = \left[\frac{\left(s_{ij}\frac{1}{\theta} + \left(1 - s_{ij}\right)\frac{1}{\eta} \right)^{-1}}{\left(s_{ij}\frac{1}{\theta} + \left(1 - s_{ij}\right)\frac{1}{\eta} \right)^{-1} + 1} \frac{1}{z_{ij}} \right]^{1-\eta} \Bigg/ \sum_{k=1}^{M_j} \left[\frac{\left(s_{kj}\frac{1}{\theta} + \left(1 - s_{kj}\right)\frac{1}{\eta} \right)^{-1}}{\left(s_{kj}\frac{1}{\theta} + \left(1 - s_{kj}\right)\frac{1}{\eta} \right)^{-1} + 1} \frac{1}{z_{kj}} \right]^{1-\eta} . \tag{8}$$

This delivers M_j equations in M_j unknowns, representing the Nash equilibrium of the market. Importantly, homotheticicity of preferences implies that this system of equations is *block recursive* in that it is independent of all aggregate variables, implying that markups can be recovered independently of aggregates.[18]

Aggregation. Aggregate variables determine prices and quantities as follows. Here it is useful to first define a sectoral markup μ_j and aggregate markup μ, as those numbers that satisfy:

$$p_{ij} = \mu_{ij}\frac{W}{z_{ij}} \quad , \quad p_j = \mu_j \frac{W}{z_j} \quad , \quad P = \mu \frac{W}{Z} \tag{9}$$

where the measures of productivity are defined as the weighted averages:

$$z_j := \left[M_j^{-1} \sum_{i=1}^{M_j} z_{ij}^{\eta-1} \right]^{\frac{1}{\eta-1}} \quad , \quad Z := \left[\int z_j^{\theta-1} \right]^{\frac{1}{\theta-1}} . \tag{10}$$

A first result is that combining these and the definitions of the price indexes we can show that the sectoral and aggregate markups are productivity weighted harmonic means:

$$\mu_j = \left[M_j^{-1} \sum_{i=1}^{M_j} \underbrace{\left(\frac{z_{ij}}{z_j} \right)^{\eta-1}}_{\zeta_{ij}} \mu_{ij}^{1-\eta} \right]^{\frac{1}{1-\eta}} \quad , \quad \mu = \left[\int \underbrace{\left(\frac{z_j}{Z} \right)^{\theta-1}}_{\zeta_j} \mu_j^{1-\theta} dj \right]^{\frac{1}{1-\theta}} . \tag{11}$$

Here the terms ζ_{ij} and ζ_j, which we use again below, are well-defined weights (between zero and one, and sum to one). Since markups are determined independently of aggregates, then so are μ and Z.

Wages. Given productivities in market j, $\boldsymbol{z}_j$, solving (8) delivers equilibrium market shares, independently of any aggregates, which in turn deliver markups (6). The relative price of final output to variable labor is then determined by the goods market clearing condition, which is simply the price index under p_{ij} given by (6). This can be inverted to obtain the relative price of labor:

$$P = \mu \frac{W}{Z} \quad \rightarrow \quad \frac{W}{P} = \frac{Z}{\mu} \tag{12}$$

These objects have intuitive interpretations. First, if all firms had markups $\mu_{ij} = 1$, which would be implied by an efficient allocation in which shadow prices of goods are equated to marginal revenue products, then under the normalization $P = 1$ the wage would be $W = Z$. Second, if all firms had the same productivity $z_{ij} = 1$, and markups $\mu_{ij} = \mu$, then the wage would be $W = 1/\mu$. A higher markup increases the relative price of goods, which with the normalization $P = 1$, decreases the wage. A higher aggregate markup can

[18] Throughout we abstract from wealth effects on labor supply. This result implies that wealth effects on labor supply will have no effect on markups. Equation (8) also shows that *sector productivity fixed effects* such that, for example, $z_{ijt} = \psi_j x_{ijt}$, will have no bearing on market shares and markups. Hence the incorporation of some sectors being larger than others on average, can be incorporated with no effect on our key moments.

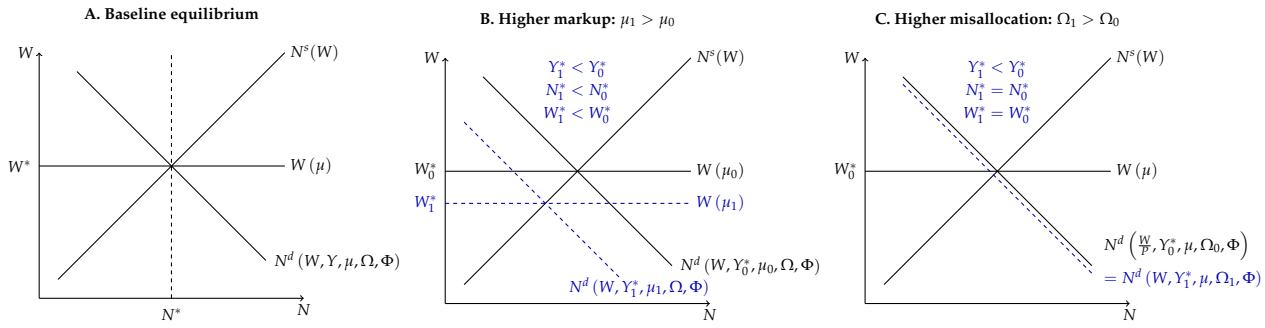

Figure 1: General equilibrium

Notes: In all cases we have normalized $P = 1$, and to save on notation set $Z = 1$. A decrease in Z would have the same effect as an increase in μ (panel B): the wage would decrease, aggregate labor supply would move along the supply curve and output would decrease such that labor demand and supply are equated. An increase in $\overline{\varphi}$ would have the same effect as a decrease in Ω (panel C): fixed costs do not enter firms' marginal conditions which pin down their price relative to the wage, and hence the equilibrium wage would remain constant. Labor supply is unchanged, to equilibrium total employment remains constant and instead labor is reallocated from production to overhead, causing output to fall.

occur due to higher firm markups μ_{ij}, or higher productivity at high markup firms, increasing the weights ζ_{ij} on high markup firms.

Labor demand. We can write the demand for labor $N^d(W, \cdot)$ in a way that similarly makes clear the wedges introduced by markups. Labor demand is the sum of the demand for labor as a variable input and the demand for labor as fixed cost, which can be written as follows, where $\Phi := \phi \int M_j dj$

$$N^d\left(\frac{W}{P}, Y, \Omega, Z, \Phi\right) = \left(\mu \frac{W/P}{Z}\right)^{-\theta} \times \frac{Y}{Z} \times \underbrace{\int \zeta_j \left(\frac{\mu_j}{\mu}\right)^{\theta} \left[\sum_{i=1}^{M_j} \zeta_{ij} \left(\frac{\mu_{ij}}{\mu_j}\right)^{\eta}\right] dj}_{\text{Wedge } \Omega \text{ due to markup heterogeneity: } \mu_{ij}} + \Phi \tag{13}$$

First, a lower wage relative to productivity moves the economy down along its labor demand curve, while a higher markup chokes off the demand for inputs, shifting the demand curve inward. The second term also shifts the labor demand curve: if Y/Z is higher, then more labor is required in production. The third term represents a wedge reflecting *misallocation*, $\Omega \geq 1$. Imposing goods market clearing on the labor demand curve implies $Y = \Omega^{-1} Z N^{Prod}$, with higher values of Ω reducing output. If all markups are identical $\Omega = 1$. If markups are heterogeneous, and positively correlated with productivity, then $\Omega > 1$, reducing output. From the perspective of labor demand, when misallocation is larger, more labor is required to produce a given level output, conditional on Z.

Equilibrium formation. Combined, we have three general equilibrium conditions. The labor supply curve, labor demand curve and the goods market clearing condition which determine $\{N, W, Y\}$:

$$N^s\left(\frac{W}{P}\right) = \overline{\varphi}\left(\frac{W}{P}\right)^{\varphi} \quad , \quad N^d\left(\frac{W}{P}, Y, \Omega, Z, \Phi\right) = \Omega\left(\mu \frac{W/P}{Z}\right)^{-\theta} \frac{Y}{Z} + \Phi \quad , \quad \frac{W}{P} = \frac{Z}{\mu}$$

Normalizing $P = 1$ and setting $Z = 1$, Figure 1 shows how this equilibrium in the labor market is formed and the economics behind comparative statics. Panel A describes how the wage is first pinned

down by goods market clearing, and then output adjusts to clear the labor market. In Panel B we consider an increase in μ. A higher markup, causes the relative price of goods to increase, which with $P = 1$ leads the wage to fall to $W_1^* < W_0^*$. As labor supply falls along its supply curve, the economy produces less goods, with production contracting until the labor market clears.[19] In Panel C, we consider a pure increase in Ω due to higher correlation of productivity and markups, keeping μ fixed. *Ceteris parabus* this expands labor demand, but does not affect the aggregate wage which depends only on Z and μ. This leaves equilibrium labor unchanged. As such, output must be lower to clear the labor market: $Y_1^* < Y_0^*$. Not shown here, we can reason through the effect of higher fixed costs Φ. A higher Φ increases labor demand, but has no effect on the wage, which is priced in terms of variable factors. With no change in the wage, equilibrium total employment is unchanged, and labor is simply reallocated from production to labor. This accommodates the required drop in output. The observed shift in the equilibrium would therefore look like Panel C.

Entry. The above delivers the general equilibrium of an economy with a given number of firms in each market $\{M_j\}_{j=0}^1$. We now consider how the equilibrium set of firms is determined. For a given set of candidate entrants in all markets, the above equilibrium delivers a wage W and output Y, which are necessary to compute firm profits. A firm will operate if:

$$\pi_{ij} = \underbrace{\left(\mu_{ij} - 1\right)\frac{W}{z_{ij}}}_{\text{Per unit profit}} \times \underbrace{y\left(p_{ij}, \mathbf{p}_{-ij}, Y\right)}_{\text{Units}} \underbrace{- W\phi}_{\text{Fixed costs}} \geq 0. \tag{14}$$

We use an equilibrium refinement device first introduced by Berry (1992). An equilibrium is selected as follows. We start by assuming all firms enter in every market ($M_j = M$ for all j). Conditional on this pattern of entry, aggregate Z is determined, and market equilibria deliver markups $\{\mu_{ij}\}_{\forall ij}$ from which we can compute the aggregate markup μ and misallocation Ω wedges. With $\{Z, \mu, \Omega\}$ pinned down, we can solve the general equilibrium for W, N and Y. This allows us to compute profits net of fixed costs (14). Of the firms with negative profits, the firm with the lowest profits exits first. Calculating profits again with one fewer firm, profits will have weakly increased.[20] This gives a new set of entrants, for which we again solve the equilibrium to obtain Z, μ, Ω, W, Y and hence profits π_{ij}, and again drop the firm with lowest, negative, profits. Continuing this way we obtain a set of firms such that, under general equilibrium prices and quantities implied by $\{Z, \mu, \Omega\}$, no firm that chooses to stay out of the market would have positive profits if they entered.[21] This algorithm delivers a *refinement of the Nash equilibrium of a simultaneous entry*

[19] In this simple case if we further set $\Phi = 0$ and $\Omega = 1$, then $Y = N = W^{\varphi} = \mu^{-\varphi}$. Note that if we allowed for wealth effects on labor supply, then the drop in output would slightly shift out the labor supply curve, reducing some of the decline in employment, and leading to a larger decline in the wage.

[20] In the computational exercise, there is a large but finite set of markets J so a large but finite set of potential entrants. Therefore a single firm exits. In the theory, with a continuum of firms, a small but positive measure of firms exits.

[21] An alternative algorithm starts with the highest productivity firm operating and considers expanding the set of entering firms down the productivity distribution until the additional entrant earns negative profits. We find that both algorithms reach the same set of firms operating in equilibrium.

game that does not require us to specify leaders and followers, as would be the case if solving for the *subgame perfect equilibrium of a sequential entry game*.[22]

2.3 Comparative statics

A key argument of this paper is that changes in primitives that relate to market structure and technology imply different outcomes for moments and aggregates. Therefore, time-series of such moments are informative of time-series of primitives, and have varying implications for aggregates.

To this extent we describe the economics behind the effects of changing the parameters that we estimate annually in our quantitative exercise below. Competition changes with the number of potential firms M, and technology changes with the size of productivity shocks σ, and the size of fixed costs ϕ. All endogenously effect the number and identity of operating firms in each market.

We study two classes of outcomes. The first are the three moments that we will use to estimate these parameters in Section 3: markups, employment reallocation rate, and the composition of costs. The second are important macroeconomic equilibrium outcomes: measured labor productivity Y/N, wage W, output Y and welfare $U(C, N)$.

Our key result is that depending on the source of the change, higher markups can be accompanied by increasing or decreasing business dynamism, increasing or decreasing measurements of costs, increasing or decreasing or flat welfare. This provides arguments for identification and sets up the counterfactual decompositions of moments and aggregates.

Business dynamism. To understand how changes in the parameters of the economy can affect business dynamism, we consider the *employment reallocation rate*. In empirical work this is often defined as total job creation and job destruction minus net job creation as a fraction of average employment over two periods (see Davis et al., 1998). In our stationary model, net job creation is zero, and aggregate employment is constant, giving the following which admits a simple decomposition:[23,24]

$$RER = \frac{JC + JD}{N} = \underbrace{\left[\frac{JC^{Inc} + JD^{Inc}}{N^{Prod}}\right]}_{\text{1. Incumbent production}} \times \underbrace{\left[1 + \frac{JC^{Ent} + JD^{Exit}}{JC^{Inc} + JD^{Inc}}\right]}_{\text{2. Composition of } JC \text{ and } JD} \times \underbrace{\left[\frac{N^{Prod}}{N^{Prod} + N^{Fixed}}\right]}_{\text{3. Share of fixed labor}}. \tag{15}$$

[22] A subtle point regarding entry is the absence of limit pricing. Limit pricing occurs in the *subgame perfect equilibrium of a sequential move game*, where a leader chooses their quantity first. The leader may choose this quantity such that a follower best-responds to this quantity by not operating. Our game features *a Nash equilibrium of a simultaneous move game*, which requires all firms' quantity decisions to be a best response to the quantity decisions of their competitors. Firms best respond to the positive quantities of their operative competitors, and the zero quantities of their inoperative competitors, and do not reason to a next step.

[23] To be precise, $JC_t := \int_j \sum_{i \in j} \max\{n_{ijt} - n_{ijt-1}), 0\} dj$ and $JD_t := \int_j \sum_{i \in j} \max\{n_{ijt-1} - n_{ijt}, 0\} dj$.

[24] An alternative moment used in the literature to capture business dynamism is the dispersion of firm growth rates: Std $(\Delta \log n_{it})$. This, however, is closely related to the reallocation rate — which is the integral under both sides of the employment change distribution. The reallocation rate can also be cleanly decomposed which we use later on.

First, reallocation occurs due to the creation and destruction of production labor jobs at incumbent firms (Inc). This depends on how shocks to firm-level marginal cost impact prices and quantities, which can be expressed in terms of the residual elasticity of demand faced by a firm and price pass-through. With linear production, employment is $n_{ijt} = y_{ijt}/z_{ijt}$. A first order approximation of a firm's employment response to a productivity shock is therefore:

$$\frac{d\log n_{ijt}}{d\log z_{ijt}} = \left(\frac{d\log y_{ijt}}{d\log p_{ijt}}\right)\left(\frac{d\log p_{ijt}}{d\log z_{ijt}}\right) - 1 = \underbrace{\varepsilon\left(s_{ijt}\right)}_{\text{Elasticity}}\ \underbrace{\chi\left(s_{ijt}\right)}_{\text{Pass-through}} - 1. \tag{16}$$

We will therefore discuss the reallocation rate of incumbents in terms of $\varepsilon(s)$ and $\chi(s)$, which are endogenous in our model. In Section 4 we show that these features rationalize the larger decline in reallocation rates at small firms relative to large firms.

Second, reallocation occurs on the extensive margin due to job creation and destruction by firms that enter (Ent) and exit $(Exit)$. The second term shows that if entering and exiting firms are larger then reallocation rates will also be higher. This will be the case if (i) the cut-off productivity for entry is higher, or (ii) firms receive large shocks that lead them to exit while being far from the cut-off in the period prior to exit.[25] Third, reallocation is mechanically lower if more labor is used in overhead, since overhead labor does not fluctuate.

Markups and composition of costs. Throughout we consider model and empirical counterparts of the economy-wide sales-weighted average of markups as in De Loecker et al. (2020),[26] which we denote $\overline{\mu}$, and the sales-weighted average of firm level fixed costs to total costs, which we denote $\overline{\Phi}$. The benefit of these measures is that they are observed in the data without having to suppose a definition of markets. These are measured identically in the model as:

$$\overline{\mu} := \int \sum_{i=1}^{M_j} \left(\frac{p_{ij}y_{ij}}{PY}\right) \mu_{ij}\, dj\,, \quad , \quad \overline{\Phi} := \int \sum_{i=1}^{M_j} \left(\frac{p_{ij}y_{ij}}{PY}\right) \frac{W\phi}{W\phi + Wn_{ij}}. \tag{17}$$

Changes in $\overline{\mu}$ can therefore be due to compositional changes — more sales accruing to higher markup firms — or within firm changes, as firms increase their markups keeping the distribution of sales fixed. In Section 4 we decompose these channels in model and data. A similar decomposition is possible for $\overline{\Phi}$, but here a reallocation of sales to larger firms would *reduce* $\overline{\Phi}$, as overhead is a smaller component of large firms'

[25] Equivalently, if a shock to z_{ij} takes a firm from below the entry threshold to far above the entry threshold, then job creation by entry, JC^{Ent}, will be larger.

[26] The recent work by Bond, Hashemi, Kaplan, and Zoch (2021), Traina (2018), Syverson (2019) and De Loecker and Syverson (2021) reminds the research community of important methodological aspects of production function estimation, which can potentially affect certain dimensions of the markup estimates. Markup estimates rely on the selection of one (or multiple) variable inputs in production and the corresponding output elasticities. The choice of variable inputs depends on data availability and varies by application, while the estimation of the output elasticities brings in all the well-known challenges that go into production function estimation. Most notably: an endogeneity bias due to unobserved productivity shocks and an omitted price bias due to unobserved input and output prices. To estimate the markups we account for these biases using the techniques suggested in the extensive literature on production function estimation. For a detailed discussion, see Appendix A in De Loecker et al. (2020).

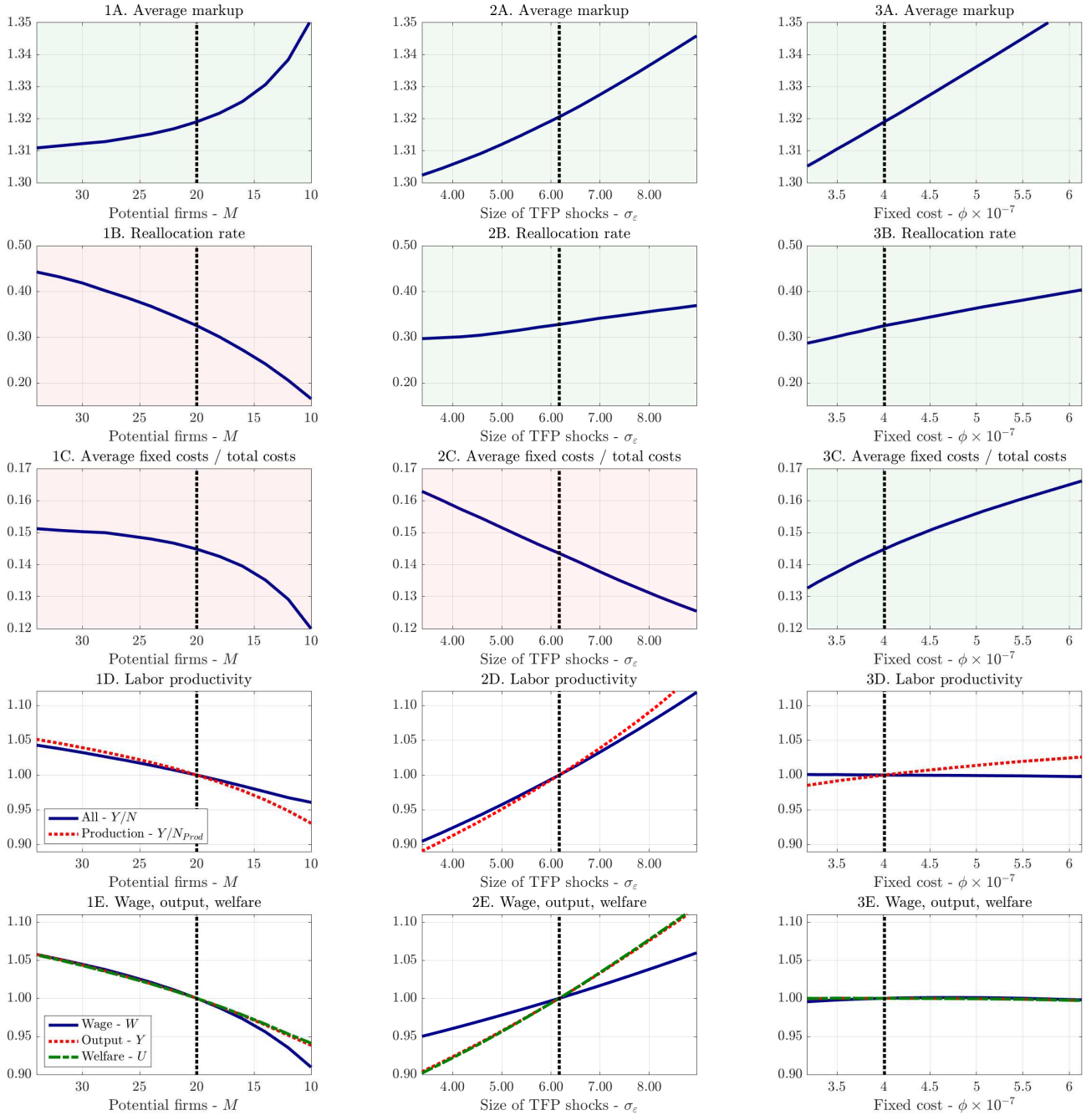

Figure 2: Equilibrium comparative static effects of changes in market structure and technology

Notes: The vertical black dashed line gives the halfway point between the minimum and maximum estimates of the relevant parameter over 1980 and 2016 from out estimates in Section 3 (see Figure 4). In each case we move only one parameter at a time, and recompute the general equilibrium of the model and the associated moments.

costs.[27]

[27] We consider alternative weights, and alternative aggregate markups series, and unless explicitly mentioned our results are robust to these alternatives.

Comparative Statics Results. All comparative static results are shown in Figure 2. The first column plots comparative statics with respect to M, the second with respect to σ, and the third with respect to ϕ. To provide context, we consider deviations of each parameter around the mid-point of the minimum and maximum estimated values over 1980 to 2016 in Section 3. This value is marked by a vertical black dashed line. Horizontal axes are ordered by comparative statics that *increase* $\overline{\mu}$ as shown in the first row of plots.

Market structure - M. We first consider the effect of a change in market structure that leads to an increase in markups: a decline in M. With fewer potential firms, holding technology fixed, the number of operating firms declines. This leads to higher markups at each firm as market shares increase, and sales are reallocated to higher markup firms, both contribute to higher $\overline{\mu}$.

It also leads to a decline in business dynamism as measured by the reallocation rate. A key contribution of the paper is to use this relationship between market power and business dynamism to understand structural changes in the US economy. As market shares increase, the residual demand elasticity $\varepsilon(s_{ij})$ faced by operating firms decreases, making quantities less responsive to changes in prices. Prices themselves are also less responsive to productivity as pass-through $\chi(s_{ij})$ is declining in market shares over the region of market shares that are obtained in equilibrium.[28] Qualitatively, this comparative static suggests a link between market power and business dynamism. Quantitatively, however, note that (i) the 25 ppt decline in the reallocation rate is quantitatively *large*, and more than the 10 ppt decline we will see in the data (Figure 3B, below), while (ii) the 4 ppt increase in markups is quantitatively *small*, and less than the 33 ppt increase we will see in the data (Figure 3A, below). Matching the data will require larger increases in markups, and smaller declines in reallocation rates.

Moreover, a decrease in M leads the composition of costs to move in the opposite direction to the data. Fewer firms are now larger on the intensive margin of production employment, reducing the level of fixed costs in total costs. Matching the data will also require a force that increases fixed costs in production.

Turning to aggregate measures of economic activity, we consider two measures of labor productivity: that which comes only from production labor (red, dashed line), and overall labor productivity for the economy as a whole (blue, solid line). As market power increases, production labor productivity Y/N_{Prod} declines. Recall that $Y = \Omega^{-1} Z N_{Prod}$. More market power leads to relatively higher markups at higher productivity firms, leading to an increase in misallocation Ω, and hence more production labor required for a given level of output. The decline in overall labor productivity, $Y/N = Y/N_{Prod} \times (N_{Prod}/N)$, is muted as with fewer firms the ratio of production to total labor (N_{Prod}/N) increases. Lower aggregate

[28] In fact $\chi(s_{ij})$ is U-shaped. With a market share around zero, markups are constant at $\mu = \eta/(\eta-1)$, and as a result pass-through $\Delta \log p / \Delta \log z = -1$. With a market share of one, markups are constant at $\mu = \theta/(\theta-1)$, and the same obtains. With interior market shares, an increase in productivity leads to an increase in market share, and an increase in the markup which yields $\Delta \log p / \Delta \log z \in (-1, 0)$. Over the relevant range of market shares obtained in the equilibrium of the estimated model, $\chi(s_{ij})$ is declining in s_{ij}, i.e. the left side of the U-shape.

productivity translates into lower wages, output and welfare (panel D).

In summary, if markups increase due only to changes in market structure, business dynamism declines, while the composition of costs moves counterfacutally, and measures of aggregate economic activity decline. We now consider technology in two parts, the effect of changes in the dispersion of shocks to variable factor productivity σ and changes to the productivity of fixed factors.

Variable cost - σ. Increasing the dispersion of productivity shocks faced by firms increases the cross-sectional dispersion of latent, or unselected, productivity: $\log z_{ijt} \sim \mathcal{N}\left(-\frac{1}{2}\frac{\sigma^2}{1-\rho^2}, \frac{\sigma^2}{1-\rho^2}\right)$. This leads to an increase in markups. As more productive firms become more productive, and less productive firms become less productive, the market shares of highly productive firms increase, increasing their markups. Also, with higher productivity, a greater share of sales is allocated to them. Both of these forces lead the sales-weighted markup $\overline{\mu}$ to increase. Consider a market with two firms. Then a mean-preserving spread leads to a higher market share of the more productive firm as well as higher markups. Eventually, as that firm's market share approaches one, it behaves as a monopolist, despite there being two competitors.

The increase in volatility increases business dynamism and reduces the ratio of fixed to total costs. First, with more volatility in productivity, the reallocation rate increases, counteracting any decrease in pass-through and lower demand elasticities of larger firms. Second, more dispersion in productivity leads the fraction of costs that are fixed to decline. Larger firms become larger on the intensive margin of production labor, causing the share of fixed costs in total costs to fall.

Contrary to the implications of higher markups via changes in market structure, an increase in markups due to changes in variable productivity leads to higher labor productivity. An increase in dispersion of productivity with endogenous entry leads to asymmetric productivity effects. Firms that otherwise were not entering have lower productivity, while firms that otherwise were entering now have higher productivity. Holding the composition of incumbents fixed, the improved right tail reallocates sales to increasingly more productive firms, increasing aggregate productivity. This reallocation reduces the market share and hence profits of previously marginal incumbents, leading some marginal incumbents to exit. This improves the selection of operating firms, further increasing aggregate productivity.

In terms of aggregates, the increase in aggregate productivity leads to higher wages, higher output and higher welfare. The effect on wages is smaller in magnitude due to higher markups dampening the outward shift in labor demand.

Fixed cost - ϕ. Finally we consider a change in fixed costs ϕ. Increasing fixed costs have a similar effect as a decrease in M on markups. With higher fixed costs, less firms operate, loosening competition and increasing the market shares and markups of the most productive firms. Similar to an increase in σ, the reallocation rate increases. This may seem puzzling, given that with fewer firms demand elasticities are

A. Comparative static		B. Moments			C. Aggregates			
		Markup	Reallocation	Fixed costs / Total costs	Y/N	Y	W	U
I. Market structure	$\downarrow M$	+	-	-	-	-	-	-
II. Technology	$\uparrow \sigma$	+	+	-	+	+	+	+
	$\uparrow \phi$	+	+	+	-	+/-	+/-	+/-
U.S. data 1980-2016		+	-	+				

Table 1: Summary of comparative static results.

lower and pass-through rates are lower. The off-setting force is job creation and job destruction by entrants, the second term in (15). Under higher fixed costs, entering and exiting firms have higher productivity and are larger when operating. As opposed to shedding a few workers when hit with a negative productivity shock, these firms now exit leading to large amounts of job reallocation. Figure D5, shows that higher ϕ leads to a larger fraction of labor being reallocated by entry and exit.

The effects of higher fixed costs on labor productivity diverge across our two measures. With higher fixed costs, operating firms are more positively selected on productivity. More goods are produced with less production labor and hence aggregate labor productivity measured in terms of production labor increases. Despite this, overall labor productivity falls, as more labor goes into overhead and less into production. On net, the efficiency gains through higher selection are eroded by higher fixed costs.

The net effect on aggregates is therefore hump-shaped, or flat relative to our previous two comparative static exercises. Efficiency gains from selection cause output, wages and welfare to increase. Efficiency losses due to higher overhead cause output, wages and welfare to decrease.

Summary. These comparative statics deliver three key results, summarized in Table 1. First, the source of increasing markups matters for welfare. Declining M reduces welfare, increasing σ increases welfare, and the off-setting selection and overhead effects of an increase in ϕ have an ambiguous effect. Second, there is no clear mapping between markups and business dynamism. A decline in M leads both markups to increase and reallocation rates to fall, while higher markups through either technology channel leads reallocation rates to increase. Third, these moments can be used to identify our desired set of primitives, are scale invariant, and avoid attempts to define markets. That they jointly put discipline on the parameters that we seek to estimate is reflected in the shading of the panels in Figure 2: green for a positive effect and red for a negative effect. Each parameter moves the moments we study in a unique pattern, which we have described in detail. Each moment is a ratio and hence scale invariant, unlike, for example, moments of the firm size distribution. We now use the data on these moments to quantify the model and understand which of these forces has been operative in the US from 1980 to 2016.

3 Quantification

We estimate the model using a combination of Compustat and Census *Business Dynamics Statistics* (BDS) data from 1980 to 2016. From the BDS we compute the annual labor reallocation rate. In order to take the model to Compustat data, we extend the model to include capital and intermediate inputs. We first describe how we extend the model, then we discuss the model and data counterparts of the moments used in the estimation. We then estimate the model annually. Throughout we normalize the productivity process to have mean 1, so we abstract from average productivity growth. Since the moments that we target are scale invariant, they are not affected by any such normalizations.

3.1 Extended model and moments

We extend the labor-only model of Section 2 so that we can map cost and markup measures from the data into the model. We treat model and data annually. This implies that the previous comparative static results apply, and the extended model only affects the mapping of the model to the data. In Appendix B we present a detailed discussion of how we extend our model to include capital and intermediates, and how we then map that model to Compustat data on sales ($Sales_{it}$), and total costs ($TotalCosts_{it}$), split into administrative expenses (SGA_{it}), costs of goods sold ($COGS_{it}$), total labor costs ($TotalLaborCosts_{it}$), and costs of capital ($CapitalCosts_{it}$).

Moments. We are interested in the following moments: the markup, the ratio of fixed costs to total costs and the reallocation rate.

1. **Markup.** The measure of markups in terms of the data and model (as discussed in Appendix B) is given by: $\mu_{it} = \alpha^{COGS}/(COGS_{it}/Sales_{it})$.

2. **Fixed costs.** For our measure of fixed to total costs we use the data to compute the ratio of total SGA_{it} to $TotalCosts_{it}$. In data and model this is:
$$\frac{SGA_{it}}{TotalCosts_{it}} = \frac{\phi}{n_{it}/\alpha^{COGS} + \phi}.$$

3. **Reallocation rate.** The reallocation rate is constructed using changes in total employment $(n_{it} + \phi)$, which gives
$$RER_t = \frac{\sum_i (n_{it+1} - n_{it})^+ + \sum_i (n_{it} - n_{it+1})^-}{\frac{1}{2}\left(\sum_i n_{it} + \sum_i n_{it+1}\right) + M_t\phi}.$$

In our underlying model discussed in detail in Appendix B, labor and intermediates are perfect substitutes in production and overhead, which leads to parameters ψ^{COGS} and ψ^{SGA} that we can choose to pin down their shares. Note that ψ^{COGS} and ψ^{SGA} do not enter into any of the above moments so in our estimation they could take on any value. For simplicity, we make the following assumption. Since we only observe total labor payments, let $\psi := \psi^{COGS} = \psi^{SGA}$. We can then measure ψ from the ratio of total labor costs to

$COGS$ and SGA:

$$\psi = \frac{TotalLaborCosts_{it}}{COGS_{it} + SGA_{it}}.$$

Since ψ^{COGS} and ψ^{SGA} do not appear anywhere in the expressions for our three moments, any value could be chosen for ψ and the moments would remain unchanged. This robustness is another advantage of the moments we have chosen.

Aggregation. When computing aggregate counterparts to the first two of these moments, we weigh by firm sales shares (the reallocation rate is already at the economy level). In Appendix B we show that this maintains the independence of the moments from ψ^{COGS} and ψ^{SGA}.

3.2 Approach

Our approach is to estimate the model every year from 1980 to 2016, where in each solution of the model aggregates are in steady-state and firms' productivity is stochastic.[29] Markups and fixed costs can be computed using data from a single period in the model. The reallocation rate requires two periods.

Estimating the model annually as a sequence of steady-states is distinct from estimating the transition of the economy from 1980 to 2016, which would require solving a complex dynamic oligopoly problem. Although this can be computed for a fixed set of parameter, it is sufficiently difficult to make estimation infeasible.

We keep the following parameters of preferences and technology constant across years:

- **Preferences.** Aggregate labor supply elasticity (φ), elasticities of substitution within and across sectors: (θ, η), discount rate (β).
- **Technology.** Production function factor demand elasticities (α^{COGS}), labor input share of labor and intermediates (ψ), depreciation rate (δ), persistence of productivity (ρ).

The parameters we vary across years were covered in our comparative statics: (i) the number of potential firms in each sector M_t, (ii) size of innovations to firm productivity σ_t, (iii) fixed labor costs ϕ_t.

This leaves the parameters θ and η, which we determine as follows. Along with θ and η, we have 3 parameters to estimate for each of 37 years: $\{M_t, \phi_t, \sigma_t\}_{t=1980}^{2016}$, giving $37 \times 3 + 2 = 113$ parameters. For these we use time-series of our 3 moments over 37 years. This leaves us short two moments for θ and η. We drop the parameters from the mid-point year 1995 from the estimation, setting them equal to their values in 1994. However we still include the 1995 moments in the estimation. This gives us $(37-1) \times 3 + 2 = 110$ parameters to estimate using $37 \times 3 = 111$ moments. The parameters θ and η are important for pinning down the average level of markups over the period, while movements in other parameters determine their path. To obtain our data for 1980 to 2016 we use data from 1978 to 2018 and apply a five year centered moving average to each of the moments.

[29] A simple alternative, for example, would be to estimate the model twice: once on pooled data from 1980 to 1990 and again on pooled data from 2010 to 2016. We instead choose to estimate the model independently for data from all 36 years of our sample.

Parameter		Value	Source
Labor Supply Elasticity	φ	0.25	Chetty et al. (2011)
Discount rate	β	0.96	Real interest rate 4% p.a.
Depreciation rate	δ	0.12	De Loecker et al. (2020)
Productivity persistence	ρ	0.90	Decker et al. (2020)
Factor share: Labor plus intermediates in variable cost	$\alpha^{COGS} = 1 - \alpha^{k}$	0.88	Compustat data
Factor share: Labor in labor plus intermediates	$\psi = \psi^{COGS} = \psi^{SGA}$	0.33	Compustat data

Table 2: Fixed parameters.

3.3 Parameters

Externally chosen. Table 2 summarizes externally chosen parameters. We set the discount factor $\beta = 0.96$ such that the real interest rate is 4 percent, and the depreciation rate $\delta = 0.12$. This gives a rental price of capital $R = (1/\beta) - (1-\delta) \approx 0.16$. For each firm in the data we can compute the α_{it}^{COGS} and ψ_{it} implied by firms' first order conditions in a given year. We take the median value within each year across firms, and then take the average across years to get a share of intermediates and labor in variable costs of $\alpha^{COGS} = 0.88$, and a share of labor in labor and intermediates of $\psi = 0.33$.[30] A high labor supply elasticity would amplify the effect of markups on aggregate quantities. We hence set the labor supply elasticity conservatively to 0.25, which is consistent with lower values of micro-estimates (Chetty et al., 2011).

We set ρ to 0.90. Given a process for TFP, the model generates a process for revenue TFP — or $TFPR$ — at the firm level. The process for $TFPR$ is less persistent. This lower persistence is due to the fact that $TFPR_i$ is proportional to the markup μ_i, and increases in productivity lead to *higher* markups. Decker et al. (2020) estimate a persistence of $TFPR$ of 0.65. A ρ of 0.90 delivers this on average over 1980 to 2016.[31,32]

Internally estimated. Figure 3 plots the moments used in estimating the model and the model's fit. The fit of the model is given by the red dashed lines, with the parameters that generate this fit given in Figure 4. The two additional parameters estimated are $\theta = 1.20$ and $\eta = 5.75$, which are consistent with alternative approaches in other papers that have studied markups and nested-CES preferences (see Atkeson and Burstein (2008), Gaubert and Itskhoki (2016), and Burstein et al. (2019)).

3.4 Model fit

With only the three degrees of freedom available, the model is able to match the data very well. Our key exercise in Section 6 will be to provide a quantitative decomposition of the changes in these moments into

[30] That is, from optimality conditions $\alpha_{it}^{COGS} = COGS_{it}/(COGS_{it} + CapitalCosts_{it})$, and $\psi_{it} = TotalLaborCosts_{it}/(COGS_{it} + SGA_{it})$. We first take medians within-t across-i of α_{it}^{COGS} and ψ_{it}. Second, we average these annual medians over time.

[31] The implied persistence of $TFPR$ in the model is estimated by OLS of $\log TFPR_{it}$ on lagged $\log TFPR_{it-1}$. In the model, the implied value for 1980 is 0.62, and the implied value for 2016 is 0.72.

[32] Decker et al. (2020) also estimate parameters of TFP processes in two different ways, however using these would not be appropriate in our context. In both cases TFP_i is obtained by deflating revenue at firm i by a common price index for sector j. This would be an incorrect procedure in our model, since firms' prices differ substantially within a sector, and in a way that is relates to firm TFP. Nonetheless, their estimates point to TFP being more persistent — with an auto-correlation of around 0.80 — than $TFPR$. We see the fact that the model endogenously generates less persistence in $TFPR$ (0.65) than TFP (0.90) as a realistic feature of the model.

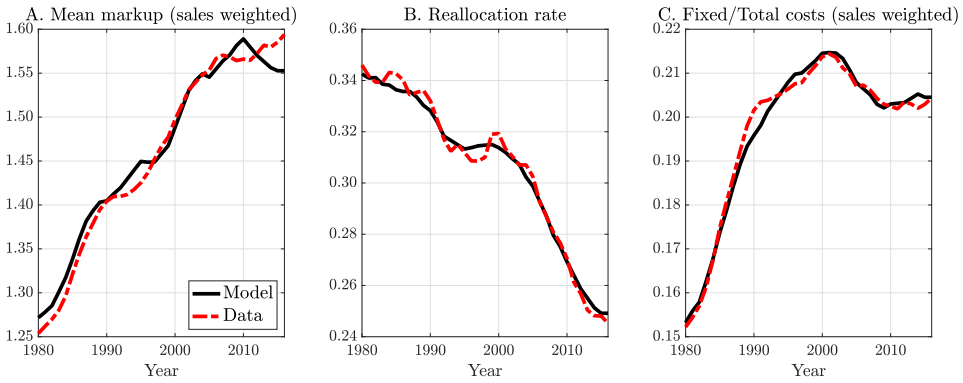

Figure 3: Model fit - Moments

Notes: Moments are computed annually, we then apply a 5 year centered moving average, which is plotted here. We target these smoothed moments in the estimation of the model, delivering the time-series of parameters which are plotted here.

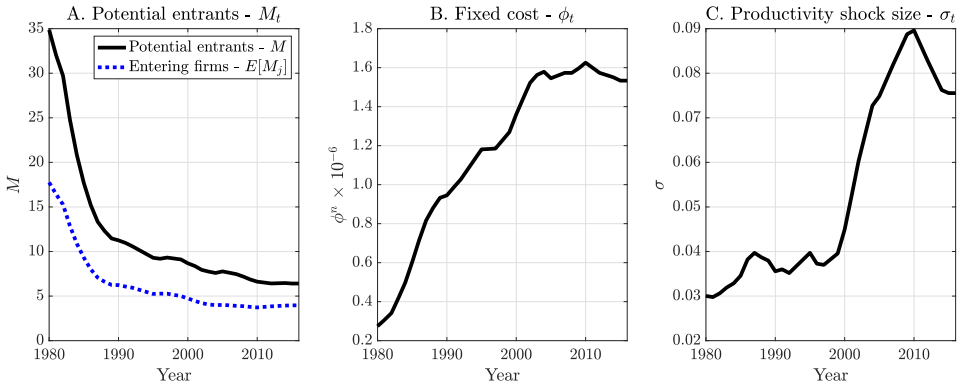

Figure 4: Parameter estimates

Notes: The actual number of firms in panel A is the unweighted average of M_j across markets.

components due to changes in each parameter. Here we provide a short description of the *quantitative* identification of the model, referring the reader back to Figure 2 which provided a qualitative argument.

To match the data both changes in technology and market structure are required: the number of potential entrants declines, productivity shocks become larger and fixed costs increase. First, consistent with our comparative static exercises a decrease in M delivers both higher markups and lower reallocation rates, two key features of the data. Second, as previously noted (Figure 2,1A-1B), a decline in M by itself would lead to a decline in reallocation rates that is quantitatively *too large* and an increase in markups that is quantitatively *too small* relative to the data. To further increase markups and dampen the decline in reallocation rates, a higher dispersion in productivity shocks is required. Third, both of these changes indirectly cause to a decline in fixed costs relative to total costs as they cause incumbents to become larger on the intensive margin of employment. This requires an off-setting increase in ϕ, which increases average fixed costs.

Appendix Figures D1 and D2 give results of an exercise that shows how the time-series for σ is identified, in particular over the increase and spike in its value post-2000. We plot the three moments under a counterfactual path for σ that smoothly joins the 2000 and 2018 values, rather than the observed path Figure 4C. Without the sharp increase in the size of productivity shocks the reallocation rate would have

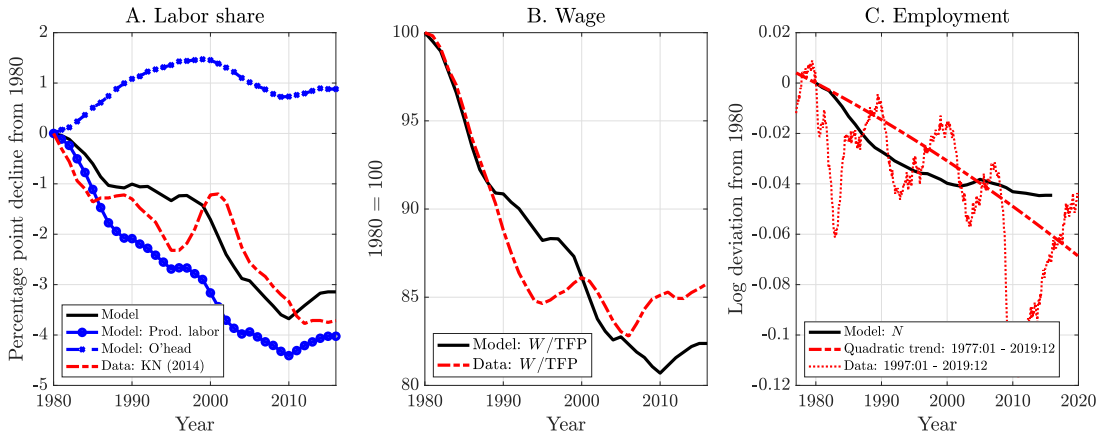

Figure 5: Validation - Aggregates

Notes: In Panel A, data is the labor share taken from Karabarbounis and Neiman (2014), in percentage point differences from 1980. In Panel B, data is average hourly earnings of production and non-supervisory employees from CPS, normalized by TFP from Penn World Tables, in level differences from 1980. In Panel C, the thin solid red-line plots the time series of the employment-population ratio for prime-aged men multiplied by the population of prime-aged men in 1980, in log differences from 1980. The red dashed line gives a quadratic fit to this data from the earliest data available (1977) through to the end of 2019, before the series declines sharply due to the Pandemic Recession.

fallen much more than in the data, and the markup would have undershot the data.

We return to a more careful decomposition of the roles of technology and market power on these moments later in the paper, as well as assessing their impact on output, wages and welfare. First we validate the model against non-targeted aggregates and cross-sectional moments over this time period.

4 Validation - Aggregates, Business Dynamism, and Markups

We study (i) the time-series implications for aggregates; (ii) the time-series of the cross-section of small and large firm reallocation rates, contributing new empirical facts; and (iii) the time-series of the decomposition of the average markup into reallocation and direct effects (De Loecker et al., 2020). These are extended over-identifying tests of the model, and confirm the empirical relevance of the model for future work.

4.1 Aggregates

Figure 5 compares the model's prediction to data on the labor share, real wage and employment.

Labor share. Data on the decline in the labor share are from Karabarbounis and Neiman (2014), plotted in red dashed lines in Figure 5A. The labor share declined by around 3.5 percentage points over this period. In the model, the aggregate labor share can be considered in two components. From the expression of the markup in equation (6) and using the firm production technology $y_{ij} = z_{ij}n_{ij}$ we obtain an expression for

the labor payment share of sales at the firm, which can then be aggregated:

$$\text{Firm:} \quad LaborShare_{ij} = \frac{Wn_{ij} + W\phi}{p_{ij}y_{ij}} = \mu_{ij}^{-1} + \frac{W\phi}{p_{ij}y_{ij}} \tag{18}$$

$$\text{Aggregate:} \quad LaborShare = \frac{\sum_{ij} Wn_{ij} + W\phi}{\sum_{ij} p_{ij}y_{ij}} = \underbrace{\sum_{ij} s_{ij}\mu_{ij}^{-1}}_{\text{Production labor}} + \underbrace{\sum_{ij} s_{ij}\frac{W\phi}{p_{ij}y_{ij}}}_{\text{Overhead labor}} \tag{19}$$

Higher markups imply lower shares of payments to labor, meanwhile higher overhead costs increase the share of payments to labor. Aggregating we have a piece due to production labor that depends on markups, and a piece that depends on overhead labor: the sales weighted average of overhead to sales. This is closely related to our targeted moment: the sales weighted average of overhead in total costs. Both components, and the overall labor share are in Figure 5A.

The model closely matches the data. Despite a large increase in the sales-weighted markup of 33 percentage points, the labor share—which depends on the sales-weighted *inverse* markup—decreases by only 4 percentage points. As pointed out by Edmond et al. (2019), the sales-weighted mean of inverse markups, or equivalently the cost-weighted markup, is relevant for aggregate efficiency as measured by the wedge ξ in the representative agent competitive factor demand condition: $W = \xi MRPL = \xi(PY)/N$. This is exactly the labor share. To the extent that the model has a smaller decline in the labor share than implied by the sales-weighted markup, it also has a smaller increase in the cost-weighted markup.

In terms of the two components, the increase in overhead production costs has a smaller effect. Nonetheless, the increase is quantitatively relevant. Absent this increase, the labor share would have fallen by 4 instead of 3 percentage points.

Wages. A key implication of our model is that as market power increases, wages decline. This occurs even if the labor market is competitive and is due to the general equilibrium effect from the decline in labor demand. Firms with market power sell less at higher prices, and therefore for a given equilibrium wage, they hire fewer workers. As a sizable fraction of firms in the economy have higher market power, this leads to a decline in aggregate output and hence a decline in the aggregate demand for labor. This general equilibrium effect in turn leads to a decline in the wage rate W, which drives a wedge between productivity and wages. As discussed above, the divergence of wages and productivity is essentially another way of thinking about the decline in the labor share, but nonetheless it is still worth describing, this time taking data on wages and productivity directly rather than taking data on the labor share as the starting point,[33]

[33]Comparing the wage in the model and the data is difficult due trend changes in total factor productivity in the data. We can think of TFP in the data as having a component due to growth from research and development and so on, and a component due to the allocation of resources across firms. We cannot separate these in the data, while our model endogenously generates the latter. We therefore treat model and data the same, dividing the wage in the data and in the model by total factor productivity. This takes growth and misallocation out of the data and misallocation out of the model. For the data, we take TFP from the Penn

and measure wages using production wages computed as average hourly earnings of production (goods and services) and non-supervisory employees from the BLS. For the model, TFP is simply output divided by employment, and we have the single wage measure W.

As expected, given the labor share, the model and data measurements of W_t/TFP_t align closely, with a 20 percentage point decline in wages relative to TFP. As firms with more market power restrict output, the demand for labor falls, which moves the economy down along its labor supply curve. This is potentially a striking insight: a decline in competitiveness in the output market causes a sharp decline in wages, even if wages are determined in competitive labor markets.

Employment. In mapping the model to the data, we construct a measure of employment that abstracts from population growth, and changes in female labor force participation over the period.[34] The trend decline in employment in the data is around 6 log points. The model generates about three quarters of this decline. Note that not all of this decline is due to the decrease in the number of firms in the economy, which happens early in our sample (Figure 4). Higher fixed costs and greater dispersion in productivity shocks also lead to higher markups, thus contracting labor demand (recall Figure 2). We implement this decomposition exactly in Section 5.

4.2 Decomposing changes in business dynamism

Declining business dynamism has been described in a number of empirical papers, and has various different attributes (see in particular Decker et al., 2020, and cites therein). Our contribution is to link declining business dynamism to the rise of market power. We focus on: a. entry and the composition of job creation and destruction; and b. labor reallocation rates in the cross-section. A number of papers have studied one or the other of these trends. We study them jointly and account for cross-sectional patterns. In each case we describe the moments in the data and model, then the mechanism that leads the model to match the data.

a. Declining entry and the shifting composition of job creation and destruction

Data. Figure 6 shows that in the Census BDS data the rate of firm entry has declined by around 4 percentage points from 1980 to 2016. At the same time, the composition of total job creation and job destruction has shifted: less creation due to entry, and less destruction due to exit. Both have declined by about 4 percentage points when measured as shares of total creation and destruction.[35]

World Table, TFP data: https://fred.stlouisfed.org/series/RTFPNAUSA632NRUG

[34] We first fix the male 'prime-age' 25 to 54 population at its level in 1980. We then apply to this the employment-population ratio of the same demographic group from 1981 to 2018. So far we have abstracted from thinking about trends and cycles since most of our data up to this point has been relatively acyclical. However there are obviously large cyclical fluctuations in employment. Therefore, we fit a quadratic trend to log employment from 1977-01 through to 2019-12. Figure 5C plots the log difference in this trend relative to 1980, along with the underlying data.

[35] See also Kehrig and Vincent (2021) for the micro evidence on the relation between business dynamism and the labor share.

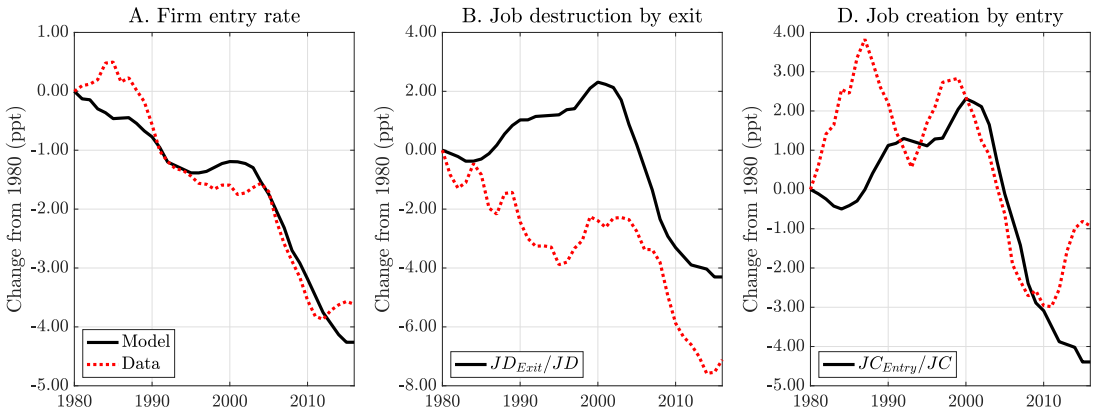

Figure 6: Validation - Entry and the composition of job creation and job destruction

Notes: Data are from the Census Business Dynamics Study. Series are first smoothed with a 7 year centered moving average and then plotted in differences from 1980. The series for job creation and job destruction are the *share* of total job creation and job destruction accounted for by firm exit and firm entry, respectively, at an annual frequency.

Model. The model accounts for these changes over time, with similar declines over the period in the entry rate of firms and shifts in the composition of job creation and job destruction. Moreover the decline is relatively gradual up to around 2000, and steeper declines since then.

Mechanism. Two effects in the model shape the extensive margin reallocation of firms (firm entry rate) and employment (job destruction by exit and job creation by entry). First, firms enter or exit following productivity shocks, and since in this granular economy markets differ according to the productivity distribution of firms, then the productivity threshold for entry varies from market to market. As the *number of potential entrant firms decreases*, the density of firms around any point, including this threshold, decreases, reducing the likelihood of one firm exiting or another firm entering. This shifts job creation and destruction in the direction of incumbents. Second, as *fixed costs increase*, the productivity threshold for operating increases. Firms that are entering and exiting become larger, increasing the fraction of job destruction by exit and job creation by entry. Endogenous market structure is, through the lens of the model, key for understanding declining entry rates and the composition of job creation and destruction. Finally, the ratio of average size of entrants to average size of incumbents remains constant, consistent with observations in Karhan et al. (2017).

b. Labor reallocation

Our economy is also consistent with new empirical facts regarding the cross-sectional properties of reallocation rates since 1980: large firms have lower reallocation rates than small firms, but reallocation at small firms fell more steeply.

Data. The aggregate employment reallocation rate R_t at date t, which we match by construction, can be decomposed into components due to small $(n < 1,000)$ and large firms $(n \geq 1,000)$, denoted by groups g:

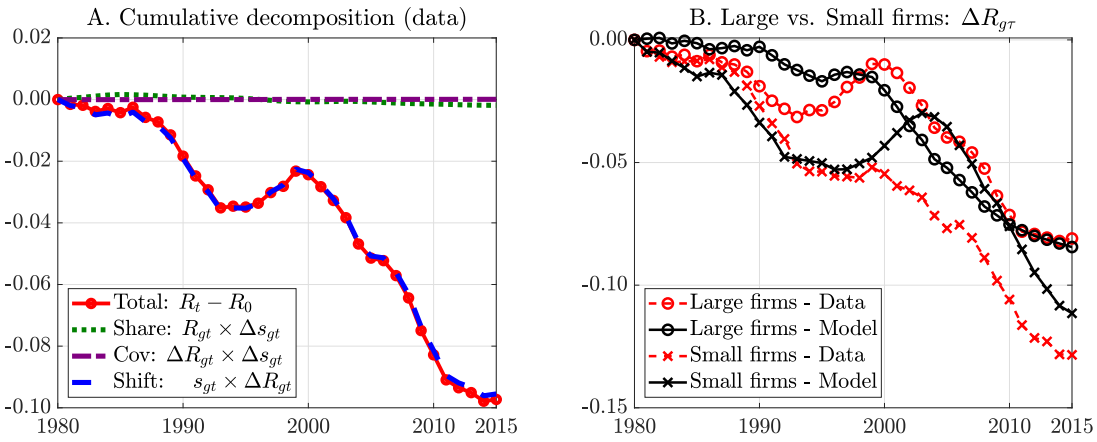

Figure 7: Validation - Decomposition of the decline in business dynamism

Notes: Panels A, B and C plot components of the decomposition of the reallocation rate using equations (20) and (21). Author's computations using Census BDS data. Large firms are defined as firms with more than 1,000 workers.

$$R_t = \frac{JC_t + JD_t}{Emp_t} = \sum_{g=1}^{G} \left(\frac{Emp_{gt}}{Emp_t}\right)\left(\frac{JC_{gt} + JD_{gt}}{Emp_{gt}}\right) = \sum_{g=1}^{G} s_{gt}^n R_{gt} \tag{20}$$

Equation (20) can be used to express the change in the reallocation rate between 1980 $(t = 1)$ and 2016 $(t = T)$ into share, shift, and covariance terms:

$$R_t - R_0 = \underbrace{\sum_{\tau=1}^{t}\sum_{g=1}^{G} R_{g\tau} \times \left[\Delta s_{g\tau}^n\right]}_{Share_t} + \underbrace{\sum_{\tau=1}^{t}\sum_{g=1}^{G} s_{g\tau} \times \left[\Delta R_{g\tau}\right]}_{Shift_t} + \underbrace{\sum_{\tau=1}^{t}\sum_{g=1}^{G} \left[\Delta s_{g\tau}^n\right] \times \left[\Delta R_{g\tau}\right]}_{Covariance_t} \tag{21}$$

Over this period the employment share of large firms increased by around 3ppt, and reallocation rates at large firms were around 10 ppt lower than at small firms. Qualitatively, therefore, the compositional shift of employment to lower reallocation rate firms via $Share_t$ could explain the decline in reallocation rates. Quantitatively, however, this accounts for barely any of the decline. Figure 7A shows that declines within size classes $(Shift_t)$ accounts for almost the entire aggregate decline. The red dashed lines in Figure 7B show that declining reallocation rates for small firms (crosses) account for most of the decline in $Shift_t$. This may come as a surprise if one thinks that changes in behavior of large firms or their increasing share of employment has been a main cause of the decline in business dynamism.

Model. The model agrees with this decomposition. First, in levels, the model generates reallocation rates that are decreasing in firm size, consistent with the data. Second, in the time-series, reallocation rates decline relatively more for small firms (Figure 7C), and decline by the same magnitudes found in the data.[36] To the best of our knowledge this is the first exercise that generates both facts, both qualitatively

[36]The model has not been calibrated to the employment distribution of firms in the economy. In defining small and large firms for this exercise we split firms by percentiles of the firm employment distribution consistent with Figure 7A. For example, in 2000 in the data, firms with over 1,000 employees account for around 65 percent of employment. We therefore compute a size cut-off in the model that we deem to be 'large' such that 'large' firms account for 65 percent of employment.

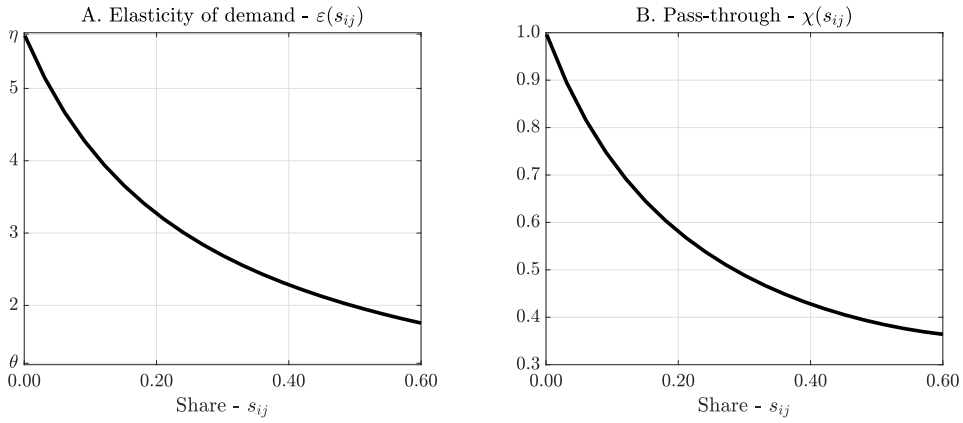

Figure 8: Elasticity of demand and pass-through by share

and quantitatively, consistent with the data.

Mechanism. Consider again the decomposition of employment changes into the demand elasticity and pass-through term, which is similar to equation (16) but expressed in terms of pass-through of marginal cost:

$$\frac{\Delta \log n_{ijt}}{\Delta \log mc_{ijt}} = \frac{\Delta \log y_{ijt}}{\Delta \log p_{ijt}} \times \frac{\Delta \log p_{ijt}}{\Delta \log mc_{ijt}} - 1 = \underbrace{\varepsilon\left(s_{ijt}\right)}_{\text{Elasticity}} \times \underbrace{\chi\left(s_{ijt}\right)}_{\text{Pass-through}} - 1. \tag{22}$$

In addition to the expression in (7) for $\varepsilon(s_{ij})$, we can derive a first order approximation for pass-through:[37]

$$\varepsilon\left(s_{ij}\right) = \left[s_{ij}\frac{1}{\theta} + \left(1 - s_{ij}\right)\frac{1}{\eta}\right]^{-1} \quad , \quad \chi(s_{ij}) \approx \frac{\left(\eta - \theta\right)\left(1 - s_{ij}\right) + \eta\left(\theta - 1\right)}{\left(\eta - \theta\right)\left(1 - s_{ij}\right)\left[1 + \left(\eta - 1\right)s_{ij}\right] + \eta\left(\theta - 1\right)} \tag{23}$$

The black solid lines in Figure 8 describe each term in (23) as a function of firms' market share.

First, consider how the model accounts for a higher *level* of reallocation rates at smaller firms. Smaller firms face both more elastic demand and pass more of any change in marginal cost through to prices, and hence quantities and employment. Second, consider how the model accounts for a larger *decline* in reallocation rates at smaller firms. Figure 8 shows that both the elasticity and pass-through terms are convex in the market share. Convexity implies that as competition declines, and revenue shares increase, both the elasticity of demand and pass-through terms decline by *more* for *smaller* firms. The result is a steeper decline in reallocation at small firms, or equivalently an observed steeper decline in the responsiveness of

[37] This is derived as in Amiti et al. (2016). As this exercise is illustrative, we only state the *direct* partial-equilibrium effect. This is also common practice in the exchange rate pass-through literature. The full expression includes sales-share weighted $\Delta \log p_{-kjt} / \Delta \log mc_{ijt}$ terms due to competitor's ($k \neq i$) equilibrium best responses to the change in firm i's marginal cost:

$$dp_{ij} = \chi(s_{ij}) dmc_{ij} + \frac{\chi(s_{ij})}{1 + \chi(s_{ij})} \sum_{k \neq i} \frac{s_{kj}}{1 - s_{ij}} dp_{kj} \quad \text{for all } i = 1, \ldots, M_j$$

These M_j equations can be solved for the full matrix of equilibrium responses Ξ_j, the diagonal of which gives the full equilibrium response of firm i to a change in its marginal cost, incorporating its response to changes in competitors' prices: $\Xi_j = (\boldsymbol{I}_j - diag(1 - \chi_{ij})\boldsymbol{W}_j)^{-1} diag(\chi_{ij})$, where $\boldsymbol{W}_j$ is a matrix that captures $\frac{s_{kj}}{1 - s_{ij}}$. In Appendix Figure D3, we plot the full response in Panel B. There is no longer closed form for the diagonal of Ξ_j. However we can compute it by solving the above system of equations for the equilibrium best-response matrix in all markets, and then take the averages by market shares of firms. The key result is that if we do so, as shown in Figure D3, the relationship is still downward sloping and convex in s_{ij}.

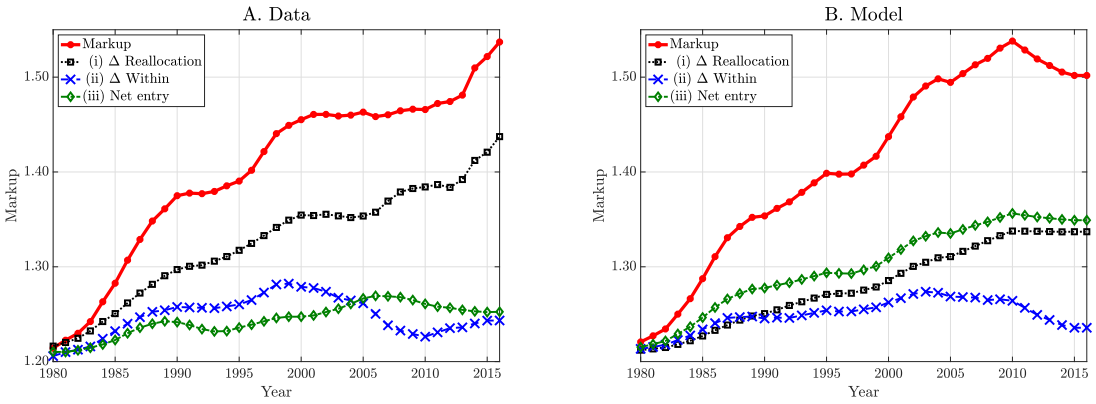

Figure 9: Validation - Decomposition of the change in markups over time

small firms to idiosyncratic shocks.

Summary. These exercises show that understanding changes in market power and its implications for pass-through can inform an understanding of changes in business dynamism. On the intensive margin (reallocation rate), extensive margin (entry rate), and in the cross-section (small vs. large) the model provides an intuitive and quantitatively accurate interpretation of the data.

4.3 Decomposing changes in markups

Next we contribute to the study of the evolution of markups by showing that the model, which matches average markups, also goes a considerable way to endogenously replicating the empirical decomposition of markups into reallocation, within-firm markup growth and changes in composition due to entry/exit.

Data. Recall that the moment that we targeted in the estimation of the model was the sales share weighted average markup $\overline{\mu}_t = \sum_i m_{it}\mu_{it}$, where as opposed to the share s_{ijt} of a firm within a sector, we use shares of sales of the entire economy m_{it}. Following De Loecker et al. (2020), we decompose $\Delta\overline{\mu}_t$ into components due to (i) changes in market shares (Δ Reallocation), (ii) changes in markups themselves (Δ Within), and (iii) the effect of Net entry as follows:

$$\Delta\overline{\mu}_t = \underbrace{\underbrace{\sum_{i\in\mathcal{I}_t\cap\mathcal{I}_{t-1}} \widetilde{\mu}_{i,t-1}\Delta m_{it}}_{\Delta \text{ Market share}} + \underbrace{\sum_{i\in\mathcal{I}_t\cap\mathcal{I}_{t-1}} \Delta\mu_{i,t}\Delta m_{it}}_{\Delta \text{ Cross term}}}_{\text{(i) } \Delta \text{ Reallocation}} + \underbrace{\sum_{i\in\mathcal{I}_t\cap\mathcal{I}_{t-1}} m_{i,t-1}\Delta\mu_{i,t}}_{\text{(ii) } \Delta \text{ Within}} + \underbrace{\sum_{i\in\mathcal{I}_t\setminus\mathcal{I}_{t-1}} \widetilde{\mu}_{it}m_{it} - \sum_{i\in\mathcal{I}_{t-1}\setminus\mathcal{I}_t} \widetilde{\mu}_{it-1}m_{it-1}}_{\text{(iii) Net entry}}. \tag{24}$$

Here $\widetilde{\mu}_{it} = \mu_{it} - \overline{\mu}_{t-1}$, and $\widetilde{\mu}_{it-1} = \mu_{it-1} - \overline{\mu}_{t-1}$, and $\mathcal{I}_t$ is the set of firms in period t. Figure 9A plots this decomposition, and shows that only around one fifth of the increase is due to the within component. The remainder is split with three fifths due to reallocation as higher markup firms capture a larger fraction of sales and another fifth due to net-entry as entering firms, on net, have higher markups.

Model. Figure 9B constructs the same decomposition using data generated from the model. This exercise should be contrasted with a similar exercise in Baqaee and Farhi (2017). There the authors take the

time-series distributions of markups from the data, treat these as exogenous wedges in a monopolistically competitive model and use the model to compute endogenous sales shares. Here we move the parameters $\{M_t, \sigma_t, \phi_t\}_{t=1980}^{2016}$, which generates an endogenous joint distributions of markups and sales shares.

Mechanism. The model generates a very similar attribution of the increase in markups to the within component, only around one fifth of the overall increase. The model then splits the remainder more evenly between reallocation and net-entry. This under-states the importance of sales reallocation with respect to the data and overstates net entry.

The decomposition shows that the rise of sales weighted markups is only partially driven by the increase in markups themselves, and much more by the reshuffling of market shares towards high markup firms. Changes in M, σ and ϕ not only effect firm level markups but are responsible for this reshuffling of the distribution of sales. Consider σ. In our model with oligopolistic competition, firms that are more productive obtain higher market shares as they can compete more effectively against firms with lower productivity. This reallocation of economic activity towards firms with high markups and high productivity has important efficiency implications and goes to the heart of the welfare implications of the rise of market power. Production by high productivity firms increases efficiency, but the variance in productivities between competing firms, exacerbated by fewer potential entrants and higher fixed costs, leads to more rent extraction and higher markups. We return to the welfare implications below.

Summary. As well as accounting quantitatively for time-series, cross-sectional and cross-sectional-time-series patterns in markups and business dynamism, we have shown that the model accounts well for key aggregate trends. This is a key step in benchmarking our main results which consist of using the model to decompose these changes in the economy — markups, business dynamism, output, wages, employment — into those parts due to changes in technology and those parts due to changes in market structure. These consistencies between model and data position the model well for the counterfactual analysis below.

5 Results - Aggregates

We now analyze the implications of changing technology and market structure for output and welfare, and ask which determinants of the model are responsible. Rather than jump to the role of each parameter in aggregates—which we provide in Figure D6—we proceed in three steps. First, we transform the aggregate equilibrium conditions into a representative agent economy with endogenous wedges. Second, we decompose output and welfare, and show that the large net decline in both output and welfare masks opposing positive and negative effects of these wedges. Third, we decompose each wedge by looking into the contribution of the primitives – technology (ϕ and σ) versus market structure (M).

None of these results rely on any local approximation of the model. In Appendix C we take a second-

order approximation of the economy to provide a variance-covariance decomposition of aggregates to better understand the distributional determinants.

5.1 Macroeconomic wedges in aggregate equilibrium conditions

To ground these exercises it is first useful to state the general equilibrium conditions of the model using the objects specified in Section 2, now in the full model with capital. The three previous conditions from the labor only model (2.2) can be written as follows with an additional set of equations that pin down output and capital given goods market clearing and labor market clearing:[38]

$$\begin{aligned}
\text{Goods market clearing:} \quad \frac{W}{P} &= \alpha \left(\frac{Z}{\mu}\right)^{\frac{1}{\alpha}} \left(\frac{1-\alpha}{R}\right)^{\frac{1-\alpha}{\alpha}} \\
\text{Total labor demand:} \quad N &= N_{Prod} + \Phi \quad , \quad W = \left(\frac{\alpha}{\mu}\right) \frac{\widetilde{Y}}{N_{Prod}} \quad , \quad \Phi := \phi \int \sum M_j dj \\
\text{Total labor supply:} \quad N &= \overline{\varphi} W^{\varphi} \\
\text{Capital demand:} \quad K &= \left(\frac{W/\alpha}{R/(1-\alpha)}\right) N_{Prod} \\
\text{Capital supply:} \quad 1 &= \beta \left[R + (1-\delta)\right] \\
\text{Output:} \quad \widetilde{Y} &= Z K^{1-\alpha} N_{Prod}^{\alpha} \quad , \quad \widetilde{Y} := \Omega Y
\end{aligned} \tag{25}$$

Given the macroeconomic wedges $\{Z, \mu, \Omega, \Phi\}$—which we can compute using only firm level productivity and markups following equations (10) and (11) from Section 2—the first three conditions can be solved for the wage W, employment N and a measure of undistorted output $\widetilde{Y}$. The remaining conditions then determine output, capital and the price R. There are numerous useful ways to express these equations. Rearranging, we can also obtain:

$$P = \mu \times \widetilde{MC} \quad , \quad \widetilde{MC} = \frac{1}{Z} \left(\frac{R}{1-\alpha}\right)^{1-\alpha} \left(\frac{W}{\alpha}\right)^{\alpha} \quad , \quad \widetilde{Y} = Z K^{1-\alpha} N_{Prod}^{\alpha} \quad , \quad Y = \left(\frac{1}{\Omega}\right) \widetilde{Y}. \tag{26}$$

Aggregate price is a markup μ on the marginal cost that one would derive from an aggregate production function with productivity Z. This distorts output, as variable factors are not priced competitively. Final output Y is further distorted by misallocation $\Omega \geq 1$. Recall that if firms have identical markups, then $\Omega = 1$, but if markups and productivity are positively correlated—as is the case in our model—then $\Omega > 1$.

5.2 Output and Welfare Decomposition

Output. Using the above equilibrium expressions we can express output in terms of total labor usage and *measured total factor productivity*:

$$Y = \underbrace{\left(\frac{1}{\widetilde{\Phi}}\right)\left(\frac{1}{\Omega}\right)\left(SZ^{*}\right)}_{\text{Measured Total Factor Productivity: } TFP} K^{1-\alpha} N^{\alpha}, \tag{27}$$

[38] In what follows we simplify the notation to $\alpha = \alpha^{COGS}$ and $1-\alpha = \alpha^{K}$.

where

$$\underbrace{\widetilde{\Phi} := \left(\frac{N}{N-\Phi}\right)^{\alpha}}_{\text{Fixed cost adjustment}} \quad , \quad \underbrace{Z^* := \left[\int \left[M^{-1}\sum_{i=1}^{M} z_{ij}^{\eta-1}\right]^{\frac{\theta-1}{\eta-1}}\right]^{\frac{1}{\theta-1}}}_{\text{Latent Productivity}} \quad , \quad \underbrace{S := \frac{Z}{Z^*}}_{\text{Selection}}.$$

This expresses output in terms of total factors and wedges which are endogenous in our model. The wedge $\widetilde{\Phi}$, which is increasing in total overhead labor Φ, reduces output through the use of total labor in overhead activities, while the misallocation wedge Ω distorts output.

Finally we express terms that depend only on firm productivity as (i) a measure of *latent productivity* (or *unselected* productivity) Z^*—which is the appropriate aggregate productivity term Z if *all* potential entrants M were to operate in all markets, that is then amplified by (ii) a measure of selection S. Latent productivity Z^* varies over time only due to changes in the parameter σ. Endogenous entry then selects higher productivity firms to operate $(S > 1)$, and improvements in selection lead S to increase. As an example, a change in the distribution of productivity via an increase in σ will show up directly as an increase in Z^* due to convexity in z_{ij}, but also indirectly through higher equilibrium S: higher productivity at top firms reduces the profits of small firms, leading some small firms to drop out and S to increase.

The decomposition in (27) does not involve the markup wedge μ. Adding the goods market clearing condition, labor supply and capital demand, includes the equilibrium effect of the markup wedge and admits a characterization of employment and output in terms of only the five wedges in the economy. For simplicity we now express terms from (25) in exact log changes, using notation $x_t = \Delta \log X_t$:

$$y_t = \underbrace{\left(\frac{1+\varphi}{\alpha}\right)\left(z_t^* + s_t\right)}_{\text{Productivity}} - \underbrace{\left(\frac{1-\alpha+\varphi}{\alpha}\right)\mu_t}_{\text{Markups}} - \underbrace{\left(\frac{1}{\alpha}\right)\widetilde{\phi}_t}_{\text{Fixed costs}} \underbrace{-\omega_t}_{\text{Misalloc.}} \quad , \quad n_t = \underbrace{\left(\frac{\varphi}{\alpha}\right)\left(z_t^* + s_t\right)}_{\text{Productivity}} - \underbrace{\left(\frac{\varphi}{\alpha}\right)\mu_t}_{\text{Markups}} \tag{28}$$

In addition to the direct effects of higher productivity $(z_t^* + s_t)$ through TFP keeping inputs fixed, higher productivity in general equilibrium increases demand for labor, which drives up the wage with elasticity φ, increasing consumption and output. Markups have a similar, but oppositely signed effect: higher markups, holding marginal costs fixed, choke off demand for labor, reducing labor income, consumption and output.[39]

Figure 10A plots the decomposition of output equation (28). Our main result is that markups and fixed costs, alone, would have lead to around a 15 percent decline in output: 8 percent from markups, and 7 percent from the rise in fixed costs.[40] While misallocation effects are small, the 15 percent decline due to markups and fixed costs is half offset by the combined increase in productivity due to innate changes in

[39] These effects are amplified by $\alpha < 1$. Since the supply of capital by households is inelastic at R, the lower is α the larger the movements in the price of labor that are required in general equilibrium.

[40] Real output in the US economy grew by 2.63 percent annually over this time, while the model implies an annual growth rate of minus 0.29 percent. This implied an off-setting trend in aggregate productivity growth of $\gamma = 2.92$ percent per year. In the model we set $\mathbb{E}[z_{ijt}] = 1$. We could include this aggregate productivity growth at rate γ by (i) setting $\mathbb{E}[z_{ijt}] = (1+\gamma)^t$, (ii) incorporating balanced growth preferences, and (iii) scaling fixed costs $\gamma^t \phi$. We could also include expansion in the number of sectors J, which we assume is constant.

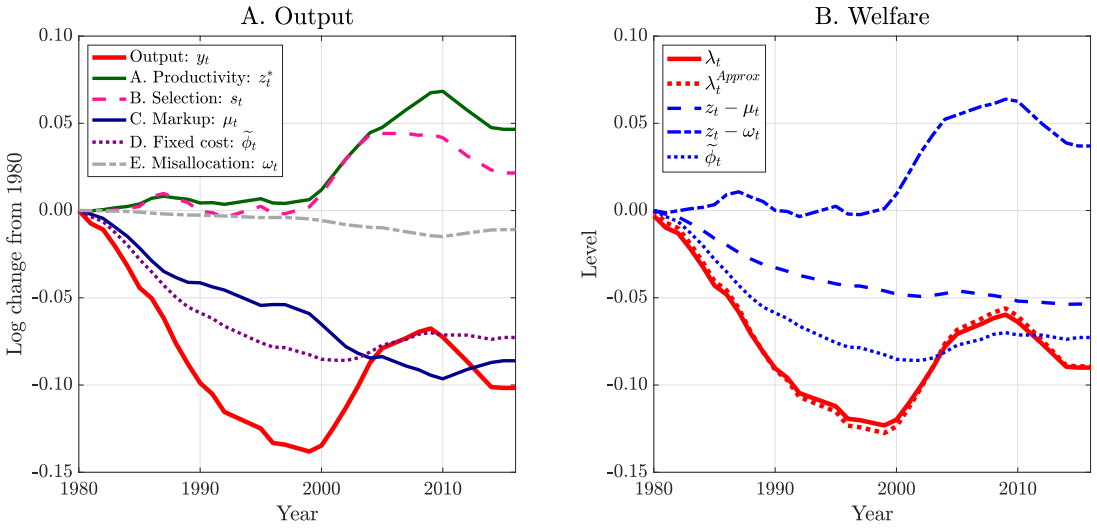

Figure 10: Decomposing output into components due to endogenous macroeconomic wedges

Notes: Panel A plots the cumulative decomposition of output according to equation (28). Panel B plots consumption equivalent welfare losses relative to 1980 λ_t, along with the decomposition of the first order approximation of these welfare losses from equation (29).

the productivity distribution (due to higher variance in the shocks) and better selection conditional on this distribution (the selection of firms that enter are of higher productivity).[41]

This is a key insight from our analysis. Output declines by 10 percent, but underneath this net decline, there is a much bigger decline that is partly offset by the increase in productivity. This shows that technological change plays a key role in the evolution of market power. Firms have become more productive incurring higher fixed costs; this has lead to fewer firms entering who do not pass on all those productivity gains to the customer, resulting in higher deadweight loss. The net effect is negative as more productive firms extract even more rents.

Welfare. We can apply a similar decomposition to welfare. We measure welfare in consumption equivalent terms, and consider the change in consumption λ_t that would be required to make the household in 1980 indifferent with respect to the period t allocation: $U((1+\lambda_t)C_{1980}, N_{1980}) = U(C_t, N_t)$. Taking a first order approximation around (C_{1980}, N_{1980}) under $\delta = 1$ and writing $x_t = \log X_t - \log X_{1980}$, the above expressions for y_t and n_t gives equation (29) which conceptualizes the effects of markups and misallocation as races against off-setting productivity effects, where $z_t = z_t^* + s_t$:

$$\lambda_t^{Approx} = y_t - ls_{1980} n_t = \underbrace{\left(\frac{1-\alpha+(1-ls_{1980})\,\varphi}{\alpha}\right)\left(z_t - \mu_t\right)}_{\text{Productivity vs. Markups}} + \underbrace{\left(z_t - \omega_t\right)}_{\text{Prod. vs. Misalloc.}} - \underbrace{\left(\frac{1}{\alpha}\right)\widetilde{\phi}_t}_{\text{Fixed costs}}, \quad ls_{1980} = \frac{W_{1980}N_{1980}}{Y_{1980}} \tag{29}$$

[41] Appendix Figure D8 plots the time-series for each of $\{z_t^*, s_t, \phi_t, \omega_t, \mu_t\}$; and Appendix Figure D9 plots the same decomposition for employment, the wage and total labor productivity Y_t/N_t, which declines by 6 percent. Total labor productivity is not the welfare relevant measure of productivity in the economy, but nonetheless is often used in empirical work. The decline in labor productivity is mostly driven by the change in composition of employment, and partly by the increase in markups, again with large off-setting effects through z_t^* and s_t.

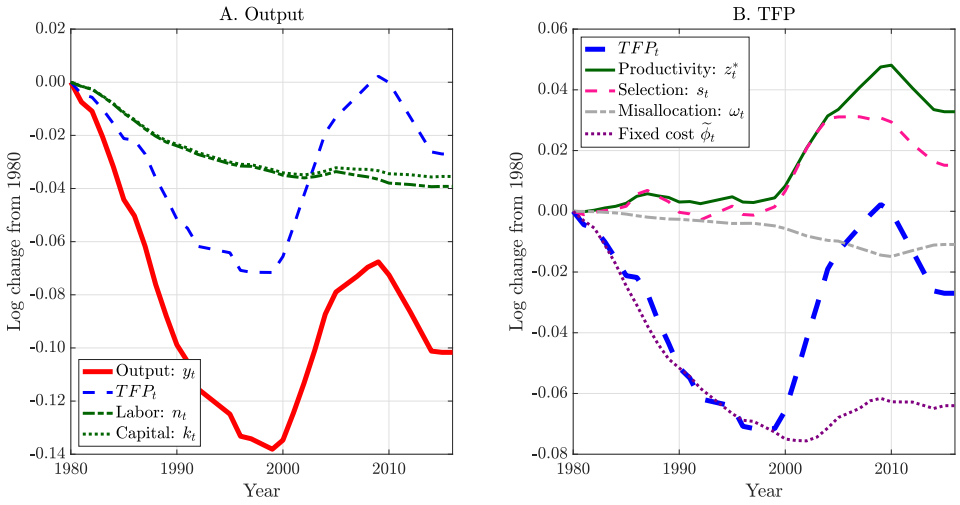

Figure 11: Decomposing output and total factor productivity

Figure 10B implements (29) and shows that welfare declines by 9 percent, about the same amount as output, and also that the first order approximation tracks the exact expression closely. While productivity effects more than offset the decline in welfare due to misallocation, the increase in markups washes out these effects, leading to a decline in welfare. Similarly there are large welfare costs associated with the change in the composition of employment. As with output, underneath the 9 percent decline in welfare, there are off-setting effects, with a large positive productivity effect on welfare.

Decomposing measured total factor productivity. We can also decompose output into the standard components. Figure 11 plots output in terms of factors and total factor productivity and the components of TFP. Panel A shows that the 10 percent decline in output from 1980 to 2016 is in nearly equal parts due to capital, labor and TFP, with TFP somewhat larger.

Panel B shows again the rich off-setting forces shaping the 6 percent decline in aggregate TFP over this period. The change in the composition of labor inputs away from variable and towards fixed factors alone would have reduced TFP by more than 6 percent, with a modest 1 percent decline due to misallocation. Off-setting these are increases in productivity after 2000 through the two channels in (27), both of which account for around a half of the increase in Z. Unselected productivity increases by over 3 percent, while there is also additional selection which further increases TFP by 2 percent.

Income effects. As an aside, note that throughout we have assumed preferences that imply no income effects on labor supply. This is most notable through equilibrium labor demand (28), which does not depend on ω_t. Since ω_t reduces output one for one, an income effect would lead labor supply to shift outward, increasing employment. Income effects would also dampen the effects of productivity and markups on output as lower output shifts labor supply out. Quantitatively we have found that our conclusions would be unchanged for reasonable values of the coefficient of relative risk aversion under separable preferences.

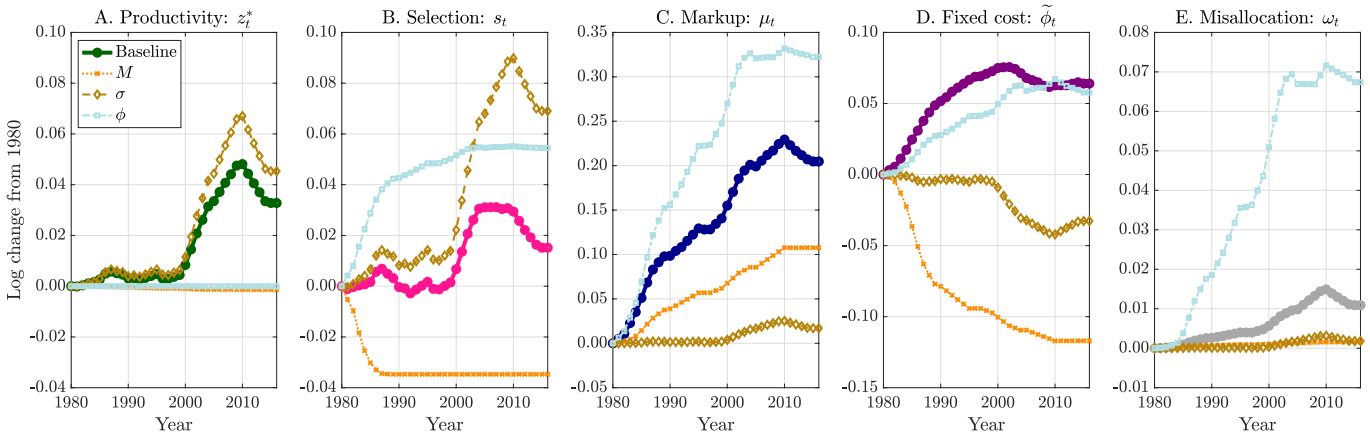

Figure 12: Effect of each parameter on the five wedges

Notes: This figure plots the effects of changing parameters independently for each wedge. For example, the orange dotted lines with square markers give the time-series of wedges implied by feeding only the estimated time-series for $\{M_t\}_{t=1980}^{2016}$ into the model, keeping ϕ_t and σ_t fixed at their estimated 1980 values.

Reduced form vs. structural decomposition. Our discussion of Figure 10 is a reduced form decomposition. If presented with sufficient data one could *measure* the reduced form wedges $\{z_t, \omega_t, \phi_t, \mu_t\}$ as residuals from a just-identified system of aggregate equations such as (25). One could then change each wedge and use the same set of conditions to solve for counterfactual paths for aggregates. This would ignore the underlying correlation structure of the reduced form wedges due to the primitive changes in the structural parameters of the economy: $\{\phi_t, M_t, \sigma_t\}$. Estimating our model over time allows us to unpack this correlation structure and account for wedges in terms of primitives, which we now turn to.

5.3 The role of market strucuture and technology in determining aggregate wedges

Now that we understand how the wedges $\{z_t^*, s_t, \mu_t, \widetilde{\phi}_t, \omega_t\}$ impact output and welfare, we can use the model to understand the contribution of each of the primitives $\{\phi_t, M_t, \sigma_t\}$ to these wedges. The primitives are the cause of changes in the economy. To do this we hold all parameters fixed at their 1980 values. We then feed in one parameter at a time, plotting the implied wedges. Figure 12 plots the wedges in panels A through E in the same order and color as Figure 10. Figure D7 provides a schematic reference to how changes in each parameter impact each wedge, which may be useful through this subsection.

Productivity (z_t^*)**.** Latent productivity z_t^* increases by around 3 percent, which by construction is almost entirely driven by the increase in dispersion of productivity from 2000 onwards. Higher dispersion increases Z_t^* due to the concavity in preferences. Recall from our discussion of Figures D1 and D2 that σ_t was identified off of the joint path of markups and reallocation rates over this period.

Selection (s_t)**.** This change to the productivity process also has positive effects through selection that magnify this direct effect on latent productivity. If the productivity cut-off for entry in a market is high enough, then increasing dispersion in productivity σ_t increases mean productivity conditional on entry.

Figure 4A showed that around half of potential firms enter in each period which is sufficiently into the right tail for increasing dispersion to increase mean productivity. Higher fixed costs ϕ_t increase the threshold for firms to enter in each market, having a substantial effect on selection. Finally the decline in the number of firms M_t, which by definition has no effect on latent productivity z_t^*, substantially reduces productivity via selection. With fewer firms, the profits associated with being an incumbent firm increase, causing the entry threshold to fall, and reducing the productivity of the marginal entrant. This negative selection effect due to changing market structure offsets all the selection gains due to mechanically increasing ϕ_t.

Markup (μ_t)**.** The aggregate markup wedge is driven by technology via increasing fixed costs, and market structure, while the dispersion in productivity has quantitatively only small effects. Higher fixed costs ϕ_t and lower competition M_t both increase markups through a similar channel: fewer incumbent firms operating. While the latter does not affect selection on productivity substantially, a higher ϕ_t does, leading to even larger increases in markups. Interestingly, when both are operating the decrease in M_t serves to dampen the effect of the increase in ϕ_t. Since each sector is granular, an increase in fixed costs may not change the number of incumbent firms if the productivity of the marginal entrant is sufficiently above the entry cutoff. As M_t declines, the density of firms near the entry cutoff declines, muting the effect of the increase in ϕ_t.

Fixed costs $(\widetilde{\phi}_t)$**.** Clearly the increase in fixed costs is driven by an increase in the parameter ϕ_t, which directly increases the fixed component of labor. The increase in the dispersion of productivity after 2000 reduces this wedge, as increased productivity for large firms increases their employment of variable labor, reducing the economy's proportional use of overhead labor. Finally, a decline in M_t mechanically has a large negative effect on the proportional use of fixed costs, since fewer firms operate. Again, these effects are dampened when the changes in market structure and technology are combined.

Misallocation (ω_t)**.** Finally, recall that the effects of misallocation on output and welfare were small. Despite this, if fixed costs alone had increased these effects would have been nearly an order of magnitude larger. Higher fixed costs induce severe misallocation as they deliver more market power to more productive firms.

Summary. Not only is understanding changes in market structure and technology important for understanding the set of macroeconomic wedges described in Figure 12 and hence aggregates, but understanding them *jointly* is also important. Changes in market structure and technology alone, can lead to severely different predictions for misallocation, markups and overall productivity net of selection.[42] We ended the last section with a comment that it *might* be difficult to discuss changes in the wedges in the economy independently due to the correlation structure in wedges induced by changes in structural parameters. Figure D11

[42]For reference, Figure D10 replicates Figure 12, removing the effects of σ and considering M and ϕ combined.

shows this is indeed the case. For example, $corr(\mu_t, \omega_t)$ is positive across all three changes in parameters, such that it is impossible to discuss the two separately. Finally, Figure D6 drops the intermediate step of macroeconomic wedges and plots the direct effect of parameters on output and welfare.

Variance-covariance decomposition. The endogenous macroeconomic wedges that determine aggregate quantities in our economy $\{z_t, \mu_t, \omega_t\}$ depend on the rich joint distribution of productivity and markups. As a final exercise Appendix C shows that, (i) the mapping from the joint distribution of productivity z_{ijt} and markups μ_{ijt} to aggregate wedges can be matched well by a second order approximation that depends on the mean, variance and covariance of markups and productivity, (ii) that, quantitatively, mean terms dominate, (iii) we then account for how primitives determine these moments. Our main result in Figure C4 is that mean terms dominate, with a small effect due to a higher variance of markups but little role for the increasing covariance between markups and productivity.

6 Results - Markups and business dynamism

From a positive perspective, what accounts for changes in measures of the average markup and business dynamism is of independent interest to economists. In this last section we return to these moments, which we matched by construction, and show that understanding the interaction between changes in market structure and technology is also important for understanding how these moments of the economy have changed over time.

Figure 13 panels B and C plots each of these moments as we feed in our estimated time-series for each parameter independently.

Increasing sales-weighted markup. Similar to the effects of each parameter on the markup wedge in Figure 12, it is also important to understand the joint effects of technology and market structure on the sales-weighted markup $\overline{\mu}$.[43] A decline in competition by itself, via a lower M_t, would raise the measured markup by about half of what is observed in the data. On the other hand, an increase in fixed costs by itself would increase the measured markup by twice what is found in the data. Together the two interact, with the decline in M_t dampening the selection effect of ϕ_t by reducing the density of firms around the exit threshold.

This presents a useful heuristic understanding of the evolution of the average markup. First, note that the effect of increasing σ_t is small, so can be ignored. Second, the number of operating firms declines by the same amount due to both higher ϕ_t and lower M_t (Panel A) Combining these observations we conclude that half of the increase in the markup is accounted for by declining competition (that which is due to M_t by itself), with the remaining half due to selection due to higher fixed costs ϕ_t.

[43] As an aside, while we match the 35 percentage point increase in the sales weighted markup (Figure 3), the aggregate markup wedge μ_t in the model increased by slightly more than half this amount: 20 percentage points.

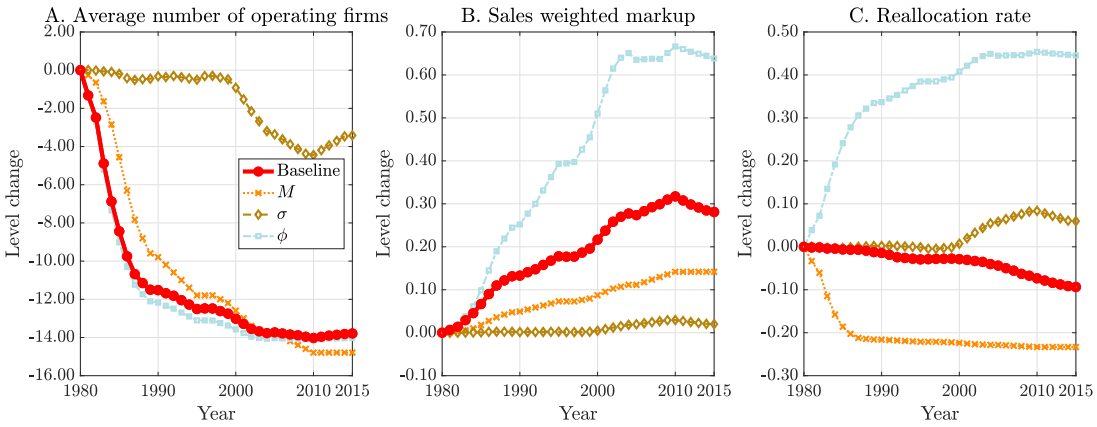

Figure 13: Decomposing the increase in sales-weighted markups and decline in business dynamism

Notes: This figure plots the effects of changing parameters independently for each wedge. For example, the orange dotted lines with square markers give the time-series of wedges implied by feeding only the estimated time-series for $\{M_t\}_{t=1980}^{2016}$ into the model, keeping ϕ_t and σ_t fixed at their estimated 1980 values.

Declining business dynamism. The observed decline in business dynamism of around 10 percentage points emerges as a balance of strongly off-setting forces. The direct decline in competition through M_t increases firms' market shares, which reduces their demand elasticities $\varepsilon(s)$ and their pass-through $\chi(s)$. These combine to reduce labor reallocation following productivity shocks. Alone this would have lead to a decline in business dynamism more than twice that observed in the data, around 22 percentage points.

Off-setting this are two forces. First, larger productivity shocks have a relatively modest effect that nonetheless offsets a quarter of the decline in dynamism due to M_t. Second, and more importantly, higher fixed costs lead to significantly higher labor reallocation. As previously discussed, higher fixed costs lead to much higher job destruction and job creation on the extensive margin. While the model matches the empirical decline in the proportion of job destruction due to exit and job creation due to entry when all parameters are changing (Figure 6), an increase in ϕ_t alone increases these terms dramatically. Once more, the interaction of declining competition and changes in technology are necessary for understanding a key aggregate trend.

7 Conclusion

Different measures suggest that market power has increased in recent decades, and this has potentially far-reaching aggregate implications for consumers, workers and households economy-wide. To assess the welfare impact of this trend, we introduce a framework that features strategic interaction between oligopolistic firms in small markets, embedded in a large economy. Entry of firms is endogenous, and both technology and market structure affect equilibrium outcomes in the product and labor market. We do this without imposing the definition of a product market, or the particular market structure. Instead, we let the moments in the data, aided by a rich micro-founded model of the macroeconomy, implicitly pin down

these objects. This is an important aspect of our analysis, and our approach underscores the notion that computing and inspecting traditional measures of industry concentration (such as HHI), is not sufficient to asses the source and impact of market power, especially in a macro context.

This framework not only allows for a quantification of the underlying sources of market power – technology and market structure – but it also provides a laboratory through we which we can evaluate the distinct role market power plays in shaping overall business dynamism in the economy. We find that both technology and market structure are necessary ingredients to explain the evolution of the major secular trends in the US economy. Technological change, predominantly through rising fixed costs, causes an increase in markups. Together with a change in market structure, through a reduction in the number of potential competitors, this leads to a decline in business dynamism as measured by lower job creation and destruction rates. It is precisely the imperfect competition in the product market that jointly determines how markups and labor demand react to changes in technology (be it in productivity or in the shift towards high fixed-cost production technologies). The decline in business dynamism is thus rooted in an incomplete pass-through of productivity shocks.

The macroeconomic implications of the rise in market power are extensive and quantitatively large. Even though the labor market is competitive, wages drop due to general equilibrium effects of an economy-wide increase in market power. With upward-sloping aggregate labor supply, our model implies a decline in labor force participation consistent with the data. This also accounts for the decline in the labor share. Taking our quantitative general equilibrium model with heterogeneous firms to the data underscores the importance of jointly allowing for technological change and changes in market structure to explain numerous secular trends in the US economy.

These profound changes result in large negative welfare effects of around 9 percent. However, this steep decline masks important opposing forces. Welfare increases substantially due to the reallocation of business towards more productive firms, but this is more than offset by these productive firms' use of their dominance to extract rents from customers. Our model and results thus unify what may seem like contradictory findings: decreasing prices yet increasing markups.

Our analysis of welfare and the decomposition of output indicate that policy implications are more subtle than myopically reducing market power. Splitting up dominant firms may decrease rent extraction, it will also destroy efficiency gains if there are scale economies. Of course, this crucially depends on whether the technology between the split up firms is rival. A simple attribution of rising market power to a weaker antitrust policy is not supported by our findings of higher efficiency of dominant firms. Instead, analyzing the impact of the dominant position of firms on product and labor market outcomes is first order.

References

ACEMOGLU, D., V. M. CARVALHO, A. OZDAGLAR, AND A. TAHBAZ-SALEHI (2012): "The network origins of aggregate fluctuations," *Econometrica*, 80, 1977–2016.

AGHION, P., A. BERGEAUD, T. BOPPART, P. J. KLENOW, AND H. LI (2019): "A theory of falling growth and rising rents," Tech. rep., National Bureau of Economic Research.

AKCIGIT, U. AND S. T. ATES (2021): "Ten Facts on Declining Business Dynamism and Lessons from Endogenous Growth Theory," *American Economic Journal: Macroeconomics*, 13, 257–98.

AMITI, M. AND S. HEISE (2021): "U.S. Market Concentration and Import Competition," Tech. rep., CEPR DP16126.

AMITI, M., O. ITSKHOKI, AND J. KONINGS (2016): "International shocks and domestic prices: how large are strategic complementarities?" NBER Working Paper 22119, National Bureau of Economic Research.

ATKESON, A. AND A. BURSTEIN (2008): "Pricing-to-Market, Trade Costs, and International Relative Prices," *American Economic Review*, 98, 1998–2031.

AUTOR, D., D. DORN, L. F. KATZ, C. PATTERSON, AND J. VAN REENEN (2017): "The Fall of the Labor Share and the Rise of Superstar Firms," Tech. rep., Centre for Economic Performance, LSE.

AZAR, J., S. BERRY, AND I. E. MARINESCU (2019): "Estimating Labor Market Power," *Available at SSRN 3456277*.

AZAR, J., I. MARINESCU, AND M. I. STEINBAUM (2017): "Labor market concentration," Tech. rep., National Bureau of Economic Research.

BAQAEE, D. R. AND E. FARHI (2017): "The Macroeconomic Impact of Microeconomic Shocks: Beyond Hulten's Theorem," Tech. rep., National Bureau of Economic Research.

BENKARD, L., A. YURUKOGLU, AND A. L. ZHANG (2021): "Concentration in Product Markets," Tech. rep., Stanford University.

BERGER, D., K. HERKENHOFF, AND S. MONGEY (2022): "Labor Market Power," *American Economic Review*, 112.

BERRY, S., M. GAYNOR, AND F. SCOTT MORTON (2019): "Do Increasing Markups Matter? Lessons from empirical industrial organization," *Journal of Economic Perspectives*, 33, 44–68.

BERRY, S., J. LEVINSOHN, AND A. PAKES (1995): "Automobile Prices in Market Equilibrium," *Econometrica*, 63, 841–890.

BERRY, S. AND P. REISS (2007): "Empirical models of entry and market structure," *Handbook of industrial organization*, 3, 1845–1886.

BERRY, S. T. (1992): "Estimation of a Model of Entry in the Airline Industry," *Econometrica*, 889–917.

BOND, S., A. HASHEMI, G. KAPLAN, AND P. ZOCH (2021): "Some Unpleasant Markup Arithmetic: Production Function Elasticities and Their Estimation from Production Data," *Journal of Monetary Economics*.

BRESNAHAN, T. F. (1982): "The oligopoly solution concept is identified," *Economics Letters*, 10, 87–92.

——— (1989): "Empirical studies of industries with market power," *Handbook of industrial organization*, 2, 1011–1057.

BURSTEIN, A., V. CARVALHO, AND B. GRASSI (2019): "Bottom-up markup fluctuations," in *2019 Meeting Papers*, Society for Economic Dynamics, 505.

CARVALHO, V. M. AND B. GRASSI (2015): "Large firm dynamics and the business cycle," Cambridge mimeo.

CARVALHO, V. M. AND A. TAHBAZ-SALEHI (2019): "Production networks: A primer," *Annual Review of Economics*, 11, 635–663.

CHARI, V. V., P. J. KEHOE, AND E. R. MCGRATTAN (2007): "Business Cycle Accounting," *Econometrica*, 75, 781–836.

CHETTY, R., A. GUREN, D. MANOLI, AND A. WEBER (2011): "Are micro and macro labor supply elasticities consistent? A review of evidence on the intensive and extensive margins," *American Economic Review*, 101, 471–75.

CHIAVARI, A. (2022): "The Macroeconomics of Rising Returns to Scale: Customer Acquisition, Markups, and Dynamism," UPF mimeo.

CROUZET, N. AND J. C. EBERLY (2019): "Understanding weak capital investment: The role of market concentration and intangibles," Tech. rep., National Bureau of Economic Research.

DAVIS, S. J., J. C. HALTIWANGER, S. SCHUH, ET AL. (1998): "Job creation and destruction," *MIT Press Books*, 1.

DE LOECKER, J., J. EECKHOUT, AND G. UNGER (2020): "The Rise of Market Power and the Macroeconomic Implications," *Quarterly Journal of Economics*, 135, 561–644.

DE LOECKER, J. AND C. SYVERSON (2021): "An industrial organization perspective on productivity," in *Handbook of Industrial Organization*, Elsevier, vol. 4, 141–223.

DE LOECKER, J. AND F. M. P. WARZYNSKI (2012): "Markups and Firm-level Export Status," *American Economic Review*, 102, 2437–2471.

DE RIDDER, M. (2019): "Market power and innovation in the intangible economy," LSE mimeo.

DECKER, R. A., J. HALTIWANGER, R. S. JARMIN, AND J. MIRANDA (2020): "Changing Business Dynamism and Productivity: Shocks vs. Responsiveness," *American Economic Review*, 110.

EDMOND, C., V. MIDRIGAN, AND D. Y. XU (2015): "Competition, Markups, and the Gains from International Trade," *American Economic Review*, 105, 3183–3221.

——— (2019): "How costly are markups?" Tech. rep., National Bureau of Economic Research.

GABAIX, X. (2011): "The granular origins of aggregate fluctuations," *Econometrica*, 79, 733–772.

GAUBERT, C. AND O. ITSKHOKI (2016): "Granular comparative advantage," Berkeley mimeo.

GRASSI, B. (2017): "IO in IO: Competition and volatility in input-output networks," *Unpublished Manuscript, Bocconi University*.

HALTIWANGER, J. (1997): "Measuring and Analyzing Aggregate Fluctuations: The Impor- tance of Building from Microeconomic Evidence," *Federal Reserve Bank St. Louis Review*, 79, 868–897.

HERSHBEIN, B., C. MACALUSO, AND C. YEH (2020): "Concentration in US local labor markets: evidence from vacancy and employment data," Tech. rep., Richmond Fed.

HOLMES, T. J. (2011): "The diffusion of Wal-Mart and economies of density," *Econometrica*, 79, 253–302.

HOPENHAYN, H. AND R. ROGERSON (1993): "Job Turnover and Policy Evaluation: A General Equilibrium Analysis," *Journal of Political Economy*, 101, 915–938.

HOPENHAYN, H. A. (1992): "Entry, exit, and firm dynamics in long run equilibrium," *Econometrica: Journal of the Econometric Society*, 1127–1150.

HOUDE, J.-F., P. NEWBERRY, AND K. SEIM (2017): "Economies of density in e-commerce: A study of amazon's fulfillment center network," Tech. rep., National Bureau of Economic Research.

JOVANOVIC, B. (1982): "Selection and the Evolution of Industry," *Econometrica*, 649–670.

KARABARBOUNIS, L. AND B. NEIMAN (2014): "The Global Decline of the Labor Share*." *Quarterly Journal of Economics*, 129.

KARHAN, F., B. PUGSLEY, AND A. SAHIN (2017): "Demographic Origins of the Startup Deficit," NBER Working Paper 25874, National Bureau of Economic Research.

KEHRIG, M. AND N. VINCENT (2021): "The micro-level anatomy of the labor share decline," *The Quarterly Journal of Economics*, 136, 1031–1087.

MONGEY, S. (2017): "Market Structure and Monetay Non-neutrality," University of Chicago mimeo.

OLMSTEAD-RUMSEY, J. (2019): "Market Concentration and the Productivity Slowdown," Mimeo.

ROSSI-HANSBERG, E., P.-D. SARTE, AND N. TRACHTER (2021): "Diverging Trends in National and Local Concentration," *NBER Macroeconomics Annual*, 35, 115–150.

SUTTON, J. (1991): *Sunk costs and market structure: Price competition, advertising, and the evolution of concentration*, MIT press.

——— (2001): *Technology and market structure: theory and history*, MIT press.

SYVERSON, C. (2019): "Macroeconomics and Market Power: Context, Implications, and Open Questions," *Journal of Economic Perspectives*, 33, 23–43.

TRAINA, J. (2018): "Is Aggregate Market Power Increasing? Production Trends using Financial Statements," Chicago Booth mimeo.

Online Appendix

This Appendix is organized as follows. Section A provides additional mathematical details and derivations. Section B provides additional details on the mapping of the model to the data. Section C provides our second order approximation of the wedges studied in the main text, and their variance-covariance decomposition. Section D provides additional Figures and Tables references in the main text.

A Derivations

A.1 Household Demand.

In the economy with CES aggregation technology, total consumption within a household C can be written as:

$$C = \left(\int_j \left(\frac{1}{J} \right)^{\frac{1}{\theta}} \left(\sum_i \left(\frac{1}{M_j} \right)^{\frac{1}{\eta}} c_{ij}^{\frac{\eta-1}{\eta}} \right)^{\frac{\eta}{\eta-1}\frac{\theta-1}{\theta}} dj \right)^{\frac{\theta}{\theta-1}} \tag{A1}$$

To derive the demand function, we first solve a maximization problem such that C is maximized with chosen c_{ij} subject to the budget constraint

$$\int_j \sum_i p_{ij} c_{ij} dj \leq Z (= WL + \Pi) \tag{A2}$$

where Z is total amount of resources. This optimization problem is equivalent to the lagarian (Maximizing the monotonic transformation of C is easier and gives the same results since C is strictly increasing in c_{ij}):

$$\mathcal{L} = \left(\int_j \left(\frac{1}{J} \right)^{\frac{1}{\theta}} \left(\sum_i \left(\frac{1}{M_j} \right)^{\frac{1}{\eta}} y_{ij}^{\frac{\eta-1}{\eta}} \right)^{\frac{\eta}{\eta-1}\frac{\theta-1}{\theta}} dj \right) - \Lambda \left(\int_j \sum_i p_{ij} y_{ij} dj - Z \right),$$

The first order condition is

$$y_{ij} = \left(\frac{p_{ij}}{p_{i'j}} \right)^{-\eta} y_{i'j}, \forall j \tag{A3}$$

Multiply both sides of (A3) by p_{ij} take the sum over i we can get

$$\begin{aligned} \sum_i p_{ij} y_{ij} &= \sum_i p_{ij}^{1-\eta} p_{i'j}^{\eta} y_{i'j}, \forall j \\ \Rightarrow Z_j &= p_{i'j}^{\eta} y_{i'j} \sum_i p_{ij}^{1-\eta}, \forall j \\ \Rightarrow y_{ij} &= \frac{Z_j p_{ij}^{-\eta}}{\sum_i p_{ij}^{1-\eta}}, \forall j \end{aligned} \tag{A4}$$

We want to derive p_j as the expenditure to buy one unit of c_j, which is $Z_j|_{c_j=1}$, and it naturally follows that

$$y_j = \left(\sum_i \left(\frac{1}{M_j}\right)^{\frac{1}{\eta}} \left(\frac{Z_j p_{ij}^{-\eta}}{\sum_i p_{ij}^{1-\eta}}\right)^{\frac{\eta-1}{\eta}}\right)^{\frac{\eta}{\eta-1}} = Z_j \left(\frac{1}{M_j}\right)^{\frac{1}{\eta-1}} \left(\sum_i p_{ij}^{1-\eta}\right)^{\frac{1}{\eta-1}}$$

$$\Rightarrow p_j = \left(\sum_i \left(\frac{1}{M_j}\right) p_{ij}^{1-\eta}\right)^{\frac{1}{1-\eta}}, \forall j$$

From (A2) we know that $\sum_i p_{ij} y_{ij} = Z_j$ and $Z_j = p_j y_j$ from the definition of p_j. We can write $\sum_i p_{ij} y_{ij} = p_j y_j$. Thus we can do similar algebra to p_j

$$\mathcal{L} = \left(\int_j \left(\frac{1}{J}\right)^{\frac{1}{\theta}} y_j^{\frac{\theta-1}{\theta}} dj\right) - \lambda \left(\int_j p_j y_j dj - Z\right),$$

and the first order condition is

$$y_j = \left(\frac{p_j}{p_{j'}}\right)^{-\theta} c_{j'}. \tag{A5}$$

We have

$$\begin{aligned} Z = \int_j p_j y_j dj = \int_j p_j \left(\frac{p_j}{p_{j'}}\right)^{-\theta} y_{j'} dj = p_{j'}^{\theta} y_{j'} \int_j p_j^{1-\theta} dj \\ \Rightarrow y_j = \frac{Z p_j^{-\theta}}{\int_j p_j^{1-\theta} dj}, \forall j \end{aligned} \tag{A6}$$

Similarly, we want to derive P as the expenditure to buy one unit of Y, which is $Z|_{Y=1}$, and it naturally follows that

$$Y = \left(\int_j \left(\frac{1}{J}\right)^{\frac{1}{\theta}} y_j^{\frac{\theta-1}{\theta}} dj\right)^{\frac{\theta}{\theta-1}} = \left(\int_j \left(\frac{1}{J}\right)^{\frac{1}{\theta}} \left(\frac{Z p_{j'}^{-\theta}}{\int_j p_j^{1-\theta} dj} dj'\right)^{\frac{\theta-1}{\theta}}\right)^{\frac{\theta}{\theta-1}} = Z \left(\frac{1}{J}\right)^{\frac{1}{\theta-1}} \int_j p_j^{1-\theta} dj^{\frac{1}{\theta-1}}$$

$$\Rightarrow P = \left(\int_j \left(\frac{1}{J}\right) p_j^{1-\theta} dj\right)^{\frac{1}{1-\theta}}$$

With y_{ij} in (A4), y_j in (A6), and $Z = p_j y_j = PY$, we can get

$$y_{ij} = \frac{Z p_{ij}^{-\eta}}{\sum_i p_{ij}^{1-\eta}} = \frac{1}{J} \frac{1}{M_j} p_{ij}^{-\eta} p_j^{\eta-\theta} P^{\theta} C,$$

A.2 Cournot Nash equilibrium

Demand within the sector is as follows

$$y_{ij} = \left(\frac{p_{ij}}{p_j}\right)^{-\eta} y_j \quad \Longrightarrow \quad p_{ij} = \left(\frac{y_{ij}}{y_j}\right)^{-1/\eta} p_j$$

Demand across sectors is

$$y_j = \left(\frac{p_j}{P}\right)^{-\theta} Y \quad \Longrightarrow \quad p_j = \left(\frac{y_j}{Y}\right)^{-1/\theta} P$$

Then the total *inverse demand function* of the firm is

$$p_{ij} = y_{ij}^{-\frac{1}{\eta}} y_j^{\frac{1}{\eta}-\frac{1}{\theta}} X$$

Using the inverse demand function, the profit function under *constant marginal cost* c_{ij}:

$$\begin{aligned}
\pi_{ij} &= y_{ij} y_{ij} - c_{ij} y_{ij} \\
&= \left[y_{ij}^{-\frac{1}{\eta}} y_j^{\frac{1}{\eta}-\frac{1}{\theta}} X \right] y_{ij} - c_{ij} y_{ij} \\
\pi_{ij} &= \underbrace{y_{ij}^{1-\frac{1}{\eta}} y_j^{\frac{1}{\eta}-\frac{1}{\theta}} X}_{r_{ij}} - c_{ij} y_{ij}
\end{aligned}$$

The first order condition is

$$0 = \left(1 - \frac{1}{\eta}\right) y_{ij}^{-\frac{1}{\eta}} y_j^{\frac{1}{\eta}-\frac{1}{\theta}} X + \left(\frac{1}{\eta} - \frac{1}{\theta}\right) y_{ij}^{1-\frac{1}{\eta}} y_j^{\frac{1}{\eta}-\frac{1}{\theta}-1} X \frac{\partial y_j}{\partial y_{ij}} - c_{ij}$$

which can be written

$$0 = \left(1 - \frac{1}{\eta}\right) \left\{ y_{ij}^{-\frac{1}{\eta}} y_j^{\frac{1}{\eta}-\frac{1}{\theta}} X \right\} + \left(\frac{1}{\eta} - \frac{1}{\theta}\right) \left\{ y_{ij}^{-\frac{1}{\eta}} y_j^{\frac{1}{\eta}-\frac{1}{\theta}} X \right\} \left[\frac{\partial y_j}{\partial y_{ij}} \frac{y_{ij}}{y_j} \right] - c_{ij}$$

Using the familiar result that $\frac{\partial y_j}{\partial y_{ij}} \frac{y_{ij}}{y_j} = s_{ij}$ under CES demand systems, and substituting back in the inverse demand function

$$0 = \left(1 - \frac{1}{\eta}\right) p_{ij} + \left(\frac{1}{\eta} - \frac{1}{\theta}\right) p_{ij} s_{ij} - c_{ij}.$$

Reearranging, we obtain

$$\begin{aligned}
p_{ij} &= \mu(s_{ij}) c_i j \\
\mu^*(s_{ij}) &= \frac{\left[s_{ij} \frac{1}{\theta} + (1 - s_{ij}) \frac{1}{\eta} \right]^{-1}}{\left[s_{ij} \frac{1}{\theta} + (1 - s_{ij}) \frac{1}{\eta} \right]^{-1} - 1}
\end{aligned}$$

A.3 Bertrand Nash equilibrium

Profits

$$\begin{aligned}
\pi_{ij} &= \left(p_{ij} - \frac{W}{z_{ij}} \right) y_{ij} \\
&= \left(p_{ij} - \frac{W}{z_{ij}} \right) \left(\frac{p_{ij}}{p_j} \right)^{-\eta} \left(\frac{p_j}{P} \right)^{-\theta} Y \\
\pi_{ij} &= \left(p_{ij} - c_{ij} \right) p_{ij}^{-\eta} p_j^{\eta-\theta} X
\end{aligned}$$

Note that

$$\frac{\partial p_j}{\partial p_{ij}} = \left[\sum_{j=1}^{M_j} p_{ij}^{1-\eta}\right]^{\frac{1}{1-\eta}-1} p_{ij}^{-\eta} = \left(\frac{p_{ij}}{p_j}\right)^{-\eta}$$

and revenue is

$$\begin{aligned} r_{ij} &= p_{ij} y_{ij} \\ &= p_{ij}\left(p_{ij}^{-\eta} p_j^{\eta-\theta} \mathrm{X}\right) \\ r_{ij} &= p_{ij}^{1-\eta} p_j^{\eta-\theta} \mathrm{X} \end{aligned}$$

which implies that revenue shares are

$$s_{ij} = \frac{r_{ij}}{\sum_k r_{kj}} = \frac{p_{ij}^{1-\eta}}{\sum_k p_{kj}^{1-\eta}} = \frac{p_{ij}^{1-\eta}}{p_j^{1-\eta}} = \frac{\partial p_j}{\partial p_{ij}} \frac{p_{ij}}{p_j}$$

We have

$$s_{ij} = \left(\frac{p_{ij}}{p_j}\right)^{1-\eta} = \frac{\partial p_{ij}}{\partial p_j} \frac{p_{ij}}{p_j}$$

The first order condition of the firm's problem is then (multiplied by -1)

$$\begin{aligned} -p_{ij}^{-\eta} p_j^{\eta-\theta} + \eta\left(p_{ij} - c_{ij}\right) p_{ij}^{-\eta-1} p_j^{\eta-\theta} - (\eta-\theta)\left(p_{ij} - c_{ij}\right) p_{ij}^{-\eta} p_j^{\eta-\theta-1} \frac{\partial p_j}{\partial p_{ij}} &= 0 \\ -p_{ij} + \eta\left(p_{ij} - c_{ij}\right) - (\eta-\theta)\left(p_{ij} - c_{ij}\right) \frac{\partial p_j}{\partial p_{ij}} \frac{p_{ij}}{p_j} &= 0 \\ -p_{ij} + \eta\left(p_{ij} - c_{ij}\right) - (\eta-\theta)\left(p_{ij} - c_{ij}\right) s_{ij} = 0, \quad \mu_{ij} &= \frac{p_{ij}}{c_{ij}} \\ p_{ij} &= \mu_{ij} c_{ij} \end{aligned}$$

where

$$\mu_{ij} = \frac{\varepsilon_{ij}}{\varepsilon_{ij} - 1}, \quad \varepsilon_{ij} = \theta s_{ij} + \left(1 - s_{ij}\right)\eta$$

A.4 Derivation of Labor Demand, equation (13)

$$N^d = \int_j \left[\sum_i^{M_j} n_{ij}\right] dj, + \int_j M_j \phi dj \tag{A7}$$

$$= \int_j \left[\sum_i^{M_j} \frac{y_{ij}}{z_{ij}}\right] dj, + \int_j M_j \phi dj \tag{A8}$$

$$= \int_j \left[\sum_i^{M_j} \frac{1}{z_{ij}} \left(\frac{p_{ij}}{p_j(p_{ij}, p_{-ij})}\right)^{-\eta} \left(\frac{p_j(p_{ij}, p_{-ij})}{P}\right)^{-\theta} Y\right] dj, + \int_j M_j \phi dj \tag{A9}$$

$$= \int_j \left[\sum_i^{M_j} \frac{1}{z_{ij}} \left(\frac{\frac{\mu_{ij} W}{z_{ij}}}{\left[\sum_{i=1}^{M_j} \left(\frac{\mu_{ij} W}{z_{ij}}\right)^{1-\eta}\right]^{\frac{1}{1-\eta}}}\right)^{-\eta} \left(\frac{\left[\sum_{i=1}^{M_j} \left(\frac{\mu_{ij} W}{z_{ij}}\right)^{1-\eta}\right]^{\frac{1}{1-\eta}}}{P}\right)^{-\theta} Y\right] dj + \int_j M_j \phi dj \tag{A10}$$

$$= Y \left(\frac{W}{P}\right)^{-\theta} \int_j \left[\sum_{i=1}^{M_j} \left(\frac{\mu_{ij}}{z_{ij}}\right)^{1-\eta}\right]^{\frac{\eta-\theta}{1-\eta}} \left[\sum_{i=1}^{M_j} \frac{1}{z_{ij}} \left(\frac{\mu_{ij}}{z_{ij}}\right)^{-\eta}\right] dj + \int_j M_j \phi dj \tag{A11}$$

We normalize the price P to be 1, from which we can compute wage W as a function of TFP z_{ij} and μ_{ij}.

$$W = \left[\int_j \frac{1}{J} \left[\sum_i \frac{1}{M_j} \left(\frac{z_{ij}}{\mu_{ij}}\right)^{\eta-1}\right]^{\frac{\theta-1}{\eta-1}} dj\right]^{\frac{1}{\theta-1}} \tag{A12}$$

This is because:

$$P = \left[\int_j \frac{1}{J} \left\{\left[\sum_i \frac{1}{M_j} p_{ij}^{1-\eta}\right]^{\frac{1}{1-\eta}}\right\}^{1-\theta} dj\right]^{\frac{1}{1-\theta}} \tag{A13}$$

$$P = \left[\int_j \frac{1}{J} \left\{\left[\sum_i \frac{1}{M_j} \left\{\mu_{ij} \frac{W}{z_{ij}}\right\}^{1-\eta}\right]^{\frac{1}{1-\eta}}\right\}^{1-\theta} dj\right]^{\frac{1}{1-\theta}} \tag{A14}$$

$$\frac{1}{P} = \left[\int_j \frac{1}{J} \left[\sum_i \frac{1}{M_j} \left(\frac{z_{ij}}{\mu_{ij}} \frac{1}{W}\right)^{\eta-1}\right]^{\frac{\theta-1}{\eta-1}} dj\right]^{\frac{1}{\theta-1}} \tag{A15}$$

$$\frac{W}{P} = \left[\int_j \frac{1}{J} \left[\sum_i \frac{1}{M_j} \left(\frac{z_{ij}}{\mu_{ij}}\right)^{\eta-1}\right]^{\frac{\theta-1}{\eta-1}} dj\right]^{\frac{1}{\theta-1}}. \tag{A16}$$

A.5 Aggregation

This section explains how we write our economy in terms of aggregates in section 2. Defining the measurement of productivity as equation 10, we start with the aggregate markups.

Markups. The definition of markups should be consistent with the nests of price indices. At the sector level, we have:

$$\begin{aligned}
p_j &= \left(\sum_i \frac{1}{M_j} p_{ij}^{1-\eta} \right)^{\frac{1}{1-\eta}} \\
&= \left[\sum_i \frac{1}{M_j} \left(\frac{\mu_{ij} W}{z_{ij}} \right)^{1-\eta} \right]^{\frac{1}{1-\eta}} \\
&= \left[\sum_i \frac{1}{M_j} \left(\frac{\mu_{ij}}{z_{ij}} \times \frac{z_j}{z_j} \right)^{1-\eta} \right]^{\frac{1}{1-\eta}} W \\
&= \left[\sum_i \frac{1}{M_j} \left(\mu_{ij} \frac{z_j}{z_{ij}} \right)^{1-\eta} \right]^{\frac{1}{1-\eta}} \frac{W}{z_j} \\
&= \left[\sum_i \frac{1}{M_j} \left(\frac{z_{ij}}{z_j} \right)^{\eta-1} \left(\frac{1}{\mu_{ij}} \right)^{\eta-1} \right]^{\frac{1}{1-\eta}} \frac{W}{z_j}
\end{aligned}$$

Noticing that W/z_j is the term indicating sectoral marginal cost, we define the bracketed term as our sector-level markups μ_j:

$$\mu_j = \left[\sum_i \frac{1}{M_j} \left(\frac{z_{ij}}{z_j} \right)^{\eta-1} \left(\frac{1}{\mu_{ij}} \right)^{\eta-1} \right]^{\frac{1}{1-\eta}},$$

or, more intuitively,

$$\frac{1}{\mu_j} = \left[\sum_i \frac{1}{M_j} \left(\frac{z_{ij}}{z_j} \right)^{\eta-1} \left(\frac{1}{\mu_{ij}} \right)^{\eta-1} \right]^{\frac{1}{\eta-1}}.$$

Similarly, we can define the economy-level aggregate as:

$$\mu = \left[\int_j \left(\frac{z_j}{Z} \right)^{\theta-1} \left(\frac{1}{\mu_j} \right)^{\theta-1} \right]^{\frac{1}{1-\theta}}.$$

Labor demand. We can also write our labor demand function in terms of above aggregates. Defining the total fixed cost as $\Phi := \int_j M_j \phi dj$, we have:

$$\begin{aligned}
N_j^d &= \sum_{i=1}^{M_j} n_{ij} + M_j\phi \\
&= \sum_{i=1}^{M_j} \left(\frac{y_{ij}}{z_{ij}}\right) + M_j\phi \\
&= \sum_{i=1}^{M_j} \left[\left(\frac{1}{J}\right)\left(\frac{1}{M_j}\right)\left(\frac{1}{z_{ij}}\right)\left(\frac{p_{ij}}{p_j}\right)^{-\eta}\left(\frac{p_j}{P}\right)^{-\theta} Y\right] + M_j\phi \\
&= \left(\frac{1}{J}\right)\sum_{i=1}^{M_j}\left[\left(\frac{1}{M_j}\right)\left(\frac{1}{z_{ij}}\right)\left(\frac{\mu_{ij}\frac{W}{z_{ij}}}{\mu_j\frac{W}{z_j}}\right)^{-\eta}\left(\frac{\mu_j\frac{W}{z_j}}{P}\right)^{-\theta} Y\right] + M_j\phi \\
&= \left(\frac{1}{J}\right)\left\{\sum_{i=1}^{M_j}\left[\left(\frac{1}{M_j}\right)\left(\frac{\mu_{ij}}{\mu_j}\right)^{-\eta}\left(\frac{z_{ij}}{z_j}\right)^{\eta-1}\left(\frac{1}{\mu_j}\right)^{\theta}\left(\frac{1}{z_j}\right)^{1-\theta}\right]\right\}\left(\frac{W}{P}\right)^{-\theta} Y + M_j\phi \\
&= \underbrace{\left[\sum_{i=1}^{M_j}\left(\frac{1}{M_j}\right)\left(\frac{z_{ij}}{z_j}\right)^{\eta-1}\left(\frac{\mu_{ij}}{\mu_j}\right)^{-\eta}\right]}_{\Omega_j}\left(\frac{1}{J}\right)\left(\mu_j\frac{W/P}{z_j}\right)^{-\theta}\frac{Y}{z_j} + M_j\phi.
\end{aligned}$$

Hence, at aggregate level, we have:

$$\begin{aligned}
N^d &= \int_j N_j^d dj \\
&= \int_j \Omega_j\left(\frac{1}{J}\right)\left(\mu_j\frac{W/P}{z_j}\right)^{-\theta}\frac{Y}{z_j}dj + \Phi \\
&= \left[\int_j \Omega_j\left(\frac{1}{J}\right)\mu_j^{-\theta}z_j^{\theta-1}dj\right]\left(\frac{W}{P}\right)^{-\theta} Y + \Phi \\
&= \underbrace{\left[\int_j \left(\frac{1}{J}\right)\Omega_j\left(\frac{\mu_j}{\mu}\right)^{-\theta}\left(\frac{z_j}{Z}\right)^{\theta-1}dj\right]}_{\Omega}\left(\mu\frac{W/P}{Z}\right)^{-\theta}\frac{Y}{Z} + \Phi.
\end{aligned}$$

Note also that at equilibrium, we have $\mu\frac{W/P}{Z} = 1$ by the definition of the markup.

B Extended model with multiple inputs

B.1 Mapping the model to the data

Accounting in the data. In Compustat we can split total costs into three components: capital costs, costs of good sold and overheard (sales and administrative expenses).[44] Costs of goods sold include labor and intermediate costs, and overhead costs also include labor and intermediates.

$$\begin{aligned} Profits_{it} &= Sales_{it} - TotalCosts_{it} && \text{(B1)} \\ TotalCosts_{it} &= CapitalCosts_{it} + \\ &\underbrace{ProdLaborCosts_{it} + ProdInterCosts_{it}}_{\text{Costs of goods sold}=COGS_{it}} + \underbrace{FixedLaborCosts_{it} + FixedInterCosts_{it}}_{\text{Overheard or Fixed costs}=SGA_{it}}. && \text{(B2)} \end{aligned}$$

In the data we observe $Sales_{it}$, $CapitalCosts_{it}$, $COGS_{it}$, SGA_{it}. We only observe a measure of total labor costs: $ProdLaborCosts_{it}+ FixedLaborCosts_{it}$. Given these constraints, we describe how we map this to the model.

Accounting in the model. We make the following assumptions to make the model consistent with (B1) and (B2): (i) the production function is constant returns to scale, (ii) labor and intermediates are perfect substitutes, and (iii) capital is used in production:

$$y_{it} = z_{it}\left(n_{it} + m_{it}\right)^{\alpha_{COGS}} k_{it}^{\alpha_k} \quad , \quad \alpha_{COGS} + \alpha_k = 1 \tag{B3}$$

Given these inputs, we have the following expression for profits in the model, which we map to the data as follows:

$$\pi_{it} = \underbrace{p_{it} z_{it}\left(n_{it} + m_{it}\right)^{\alpha_{COGS}} k_{it}^{\alpha_k}}_{Sales_{it}} - \underbrace{R_t k_{it}}_{CapitalCosts_{it}} - \underbrace{\left[P_t^m m_{it} + W_t n_{it}\right]}_{COGS_{it}} - \underbrace{\left[P_t^m \phi^m + W_t \phi\right]}_{SGA_{it}}. \tag{B4}$$

We have assumed that the firm faces the same prices for intermediates and labor regardless of whether they are used in production or in overhead. In terms of the economics, we show that the model analyzed so far with only labor remains appropriate, and how moments computed from the available data map into the model.

Optimality. Since they are perfect substitutes, the firm will be indifferent between labor and intermediates in production. We therefore assume that their shares are the same across firms, and define parameters ψ^{COGS} and ψ^{SGA}:

$$\psi^{COGS} := \frac{n_{it}}{n_{it} + m_{it}} \quad , \quad \psi^{SGA} := \frac{\phi}{\phi + \phi^m}.$$

[44] We follow De Loecker et al. (2020) for the data construction.

The first order conditions of the variable cost minimization problem give expressions for the markup and marginal cost, and deliver the result that ψ^{COGS} is also equal to the cost share of labor in $COGS_{it}$:

$$\mu_{it} := \frac{p_{it}}{mc_{it}} = \frac{\alpha_{COGS}}{COGS_{it}/p_{it}y_{it}} \quad , \quad mc_{it} = \frac{1}{z_{it}} \underbrace{\left(\frac{W_t}{\alpha_{COGS}}\right)^{\alpha_{COGS}} \left(\frac{R_t}{\alpha_k}\right)^{\alpha_k}}_{\text{Aggregate marginal cost: } MC_t} \quad , \quad \psi^{COGS} = \frac{W_t n_{it}}{W_t n_{it} + P_t^m m_{it}}.$$

Combining factor demands for intermediates and capital, we can write the production function and total variable costs in terms of only labor:

$$y_{it} = \widetilde{z}_{it} n_{it} \quad , \quad \widetilde{z}_{it} = \frac{1}{\psi^{COGS}\alpha^{COGS}} \left(\frac{W_t}{MC_t}\right) z_{it} \quad , \quad \underbrace{R_t k_{it} + W_t n_{it} + P_t^m m_{it}}_{\text{Total variable cost}} = \underbrace{\frac{W_t}{\psi^{COGS}\alpha^{COGS}}}_{:=\widetilde{W}_t} \times n_{it}$$

Re-writing the firms' problem in terms of these objects gives the following expression for profits:

$$\pi_{it} = p_{it}\widetilde{z}_{it} n_{it} - \widetilde{W}_t n_{it} - \widetilde{W}_t \left(\alpha^{COGS}\phi\right).$$

This delivers the following optimal price, where marginal cost $\widetilde{mc}_{it} = mc_{it}$ is consistent with the above:

$$\mu_{it} = \frac{p_{it}}{\widetilde{mc}_{it}} \quad , \quad \widetilde{mc}_{it} = \frac{\widetilde{W}_t}{\widetilde{z}_{it}} \quad , \quad \mu_{it} = \frac{\varepsilon_{it}}{\varepsilon_{it}+1} \quad , \quad \varepsilon_{it} = \left[\frac{1}{\theta} s_{it} + \frac{1}{\eta}\left(1 - s_{it}\right)\right]^{-1}. \tag{B5}$$

B.2 Aggregation

This section explains how we extend our aggregated notation from single-input economy into the multi-input one with the following production function where the intermediates are already substituted:

$$y_{ij} = z_{ij} \left(\frac{1}{\psi^{COGS}}\right)^{\alpha_{COGS}} \left(\frac{k_{ij}}{n_{ij}}\right)^{\alpha_K} n_{ij}.$$

We will derive all the equilibrium conditions mentioned in equation 25.

Optimality. By solving the cost minimization problem, we obtain firms' optimal combination of labor and capital:

$$\frac{k_{ij}}{n_{ij}} = \frac{1}{\psi^{COGS}} \left(\frac{W/\alpha^{COGS}}{R/\alpha^K}\right).$$

The solution also gives us the marginal cost for production:

$$mc_{ij} = \frac{1}{z_{ij}} \left(\frac{W}{\alpha^{COGS}}\right)^{\alpha^{COGS}} \left(\frac{R}{\alpha^K}\right)^{\alpha^K}.$$

Notice that this ratio is independent of firms' characteristics, which allows us to extend this optimality condition to aggregates:

$$\frac{K}{N_{prod}} = \frac{1}{\psi^{COGS}} \left(\frac{W/\alpha^{COGS}}{R/\alpha^K}\right).$$

Production technology. We first derive the production function in *aggregate* level. To do so, we aggregate production labor:

$$\begin{aligned} N_{prod}^{d} &= \int_j \sum_{i=1}^{M_j} n_{ij} dj \\ &= \left(\frac{k_{ij}}{n_{ij}}\right)^{-\alpha^K} \left(\psi^{COGS}\right)^{\alpha^{COGS}} \int_j \sum_{i=1}^{M_j} \left(\frac{y_{ij}}{z_{ij}}\right) dj \\ &= \left(\frac{K}{N_{prod}}\right)^{-\alpha^K} \left(\psi^{COGS}\right)^{\alpha^{COGS}} \Omega \frac{Y}{Z} \end{aligned}$$

By rearranging, we get:

$$\begin{aligned} \Omega Y &= \left(\frac{1}{\psi^{COGS}}\right)^{\alpha^{COGS}} Z \left(\frac{K}{N_{prod}}\right)^{\alpha^K} N_{prod} \\ \text{Output:} \quad \widetilde{Y} &= \left(\frac{1}{\psi^{COGS}}\right)^{\alpha^{COGS}} Z K^{\alpha^K} N_{prod}^{\alpha^{COGS}}, \quad \widetilde{Y} := \Omega Y \end{aligned} \tag{B6}$$

Goods market clearing. We then investigate the goods market clearing condition given the normalized aggregated price P:

$$\begin{aligned} 1 = P &= \left[\int_0^1 \left(\sum_i \frac{1}{M_j} p_{ij}^{1-\eta}\right)^{\frac{1-\theta}{1-\eta}} dj\right]^{\frac{1}{1-\theta}} \\ &= \left[\int_0^1 \left(\sum_i \frac{1}{M_j} \left(\mu_{ij} mc_{ij}\right)^{1-\eta}\right)^{\frac{1-\theta}{1-\eta}} dj\right]^{\frac{1}{1-\theta}} \\ &= \left(\frac{W}{\alpha^{COGS}}\right)^{\alpha^{COGS}} \left(\frac{R}{\alpha^K}\right)^{\alpha^K} \left[\int_0^1 \left(\sum_i \frac{1}{M_j} \left(\frac{\mu_{ij}}{z_{ij}}\right)^{1-\eta}\right)^{\frac{1-\theta}{1-\eta}} dj\right]^{\frac{1}{1-\theta}} \\ &= \left(\frac{W}{\alpha^{COGS}}\right)^{\alpha^{COGS}} \left(\frac{R}{\alpha^K}\right)^{\alpha^K} \frac{\mu}{Z}, \end{aligned}$$

which can be rewritten as:

$$\text{Goods market clearing:} \quad W = \alpha^{COGS} \left(\frac{\alpha^K}{R}\right)^{\frac{\alpha^K}{\alpha^{COGS}}} \left(\frac{Z}{\mu}\right)^{\frac{1}{\alpha^{COGS}}}. \tag{B7}$$

Labor markets clearing. We already have the labor supply function in the aggregate level:

$$\text{Total labor supply:} \quad N = \overline{\varphi} W^{\varphi}. \tag{B8}$$

On the demand side, we can express the labor demand in terms of factor payment shares from the goods market clearing condition:

$$
\begin{aligned}
P &= \left(\frac{W}{\alpha^{COGS}}\right)^{\alpha^{COGS}} \left(\frac{R}{\alpha^{K}}\right)^{\alpha^{K}} \frac{\mu}{Z} \\
1 &= \frac{P}{\mu} Z \left[\left(\frac{W}{\alpha^{COGS}}\right)^{-\alpha^{COGS}} \left(\frac{R}{\alpha^{K}}\right)^{-\alpha^{K}}\right] \\
1 &= \frac{P}{\mu} Z \left[\left(\frac{W}{\alpha^{COGS}}\right)^{-1} \left(\frac{W/\alpha^{COGS}}{R/\alpha^{K}}\right)^{\alpha^{K}}\right] \\
WN_{prod} &= \alpha^{COGS} \frac{P}{\mu} Z \left(\frac{K}{N_{prod}}\right)^{\alpha^{K}} N_{prod}
\end{aligned}
$$

Notice that the production function is embedded in the RHS:

$$
\begin{aligned}
WN_{prod} &= \alpha^{COGS} \frac{P}{\mu} Z K^{\alpha^{K}} N_{prod}^{\alpha^{COGS}} \\
WN_{prod} &= \alpha^{COGS} \left(\psi^{COGS}\right)^{\alpha^{COGS}} \frac{P}{\mu} \widetilde{Y} \\
N_{prod} &= \alpha^{COGS} \left(\psi^{COGS}\right)^{\alpha^{COGS}} \frac{\widetilde{Y}}{\mu W} \\
\text{Total labor demand:} \quad N &= \alpha^{COGS} \left(\psi^{COGS}\right)^{\alpha^{COGS}} \frac{\widetilde{Y}}{\mu W} + \Phi
\end{aligned} \tag{B9}
$$

Capital markets clearing. Moreover, the supply of capital is elastic at capital price R determined by the household Euler equation:

$$
\text{Capital supply:} \quad 1 = \beta[R + (1 - \delta)]. \tag{B10}
$$

Firm optimality gives aggregate capital demand:

$$
\text{Capital demand:} \quad K = \frac{1}{\psi^{COGS}} \left(\frac{W/\alpha^{COGS}}{R/\alpha^{K}}\right) N_{prod}. \tag{B11}
$$

B.3 Output and welfare decomposition

Output decomposition. The aggregation system enables us to express all the aggregates in terms of our wedges $\{Z^*, \Omega, S, \mu, \widetilde{\Phi}\}$, where $Z = SZ^*$:

$$W = \psi^{COGS}\alpha^{COGS}\left(\frac{\alpha^K}{R}\right)^{\frac{\alpha^K}{\alpha^{COGS}}}\left(\frac{Z^*S}{\mu}\right)^{\frac{1}{\alpha^{COGS}}}$$

$$N = \overline{\varphi}\left[\psi^{COGS}\alpha^{COGS}\left(\frac{\alpha^K}{R}\right)^{\frac{\alpha^K}{\alpha^{COGS}}}\right]^{\varphi}\left(\frac{Z^*S}{\mu}\right)^{\frac{\varphi}{\alpha^{COGS}}}$$

$$Y = \overline{\varphi}\left(\psi^{COGS}\alpha^{COGS}\right)^{\varphi}\left(\frac{\alpha^K}{R}\right)^{\frac{\alpha^K}{\alpha^{COGS}}(1+\varphi)}(Z^*S)^{\frac{1+\varphi}{\alpha^{COGS}}}\,\Omega^{-1}\mu^{-\frac{\alpha^K+\varphi}{\alpha^{COGS}}}\widetilde{\Phi}^{-\frac{1}{\alpha^{COGS}}}$$

Hence, we get:

$$y_t = \frac{1+\varphi}{\alpha^{COGS}}(z_t^* - s_t) - \left(\frac{\alpha^K+\varphi}{\alpha^{COGS}}\right)\Delta\log\mu_t - \frac{1}{\alpha^{COGS}}\widetilde{\phi}_t - \omega_t \tag{B12}$$

$$n_t = \frac{\varphi}{\alpha^{COGS}}(z_t^* - s_t) - \frac{\varphi}{\alpha^{COGS}}\Delta\log\mu_t \tag{B13}$$

Welfare decomposition. We measure welfare in consumption equivalent terms, which makes the household in 1980 indifferent with respect to the period t allocation:

$$U\left((1+\lambda_t)C_0, N_0\right) = U\left(C_t, N_t\right)$$

$$U\left(C_0,N_0\right) + U_c\left(C_0,N_0\right)\lambda_t C_0 \approx U\left(C_0,N_0\right) + U_c\left(C_0,N_0\right)\left(C_t - C_0\right) + U_n\left(C_0,N_0\right)\left(N_t - N_0\right)$$

$$U_c\left(C_0,N_0\right)\lambda_t C_0 = U_c\left(C_0,N_0\right)\left(C_t - C_0\right) + U_n\left(C_0,N_0\right)\left(N_t - N_0\right)$$

which gives us an approximation for λ_t:

$$\lambda_t = \left(\frac{C_t - C_0}{C_0}\right) + \frac{U_n\left(C_0,N_0\right)}{U_c\left(C_0,N_0\right)}\frac{N_0}{C_0}\left(\frac{N_t - N_0}{N_0}\right)$$

$$\lambda_t = \left(\frac{C_t - C_0}{C_0}\right) - \underbrace{\left(\frac{W_0}{P_0}\frac{N_0}{C_0}\right)}_{\text{Labor share}}\left(\frac{N_t - N_0}{N_0}\right)$$

$$\lambda_t = \Delta\log C_t - ls_0\cdot\Delta\log N_t$$

$$\lambda_t = \left[\frac{\alpha^K + (1-ls_0)\varphi}{\alpha^{COGS}}\right](z_t - \mu_t) + (z_t - \omega_t) - \frac{1}{\alpha^{COGS}}\widetilde{\phi}_t \tag{B14}$$

B.4 Variance-covariance decomposition

We exploit a second-order Taylor approximation to decompose aggregates into moments of the underlying distribution of markups and productivity. Here, we decompose productivity $\widetilde{Z}_t$ as an example, and all other decomposition can be carried out in the same approach.

Start with the aggregate productivity defined in a single-nest approximation:

$$\widetilde{Z}_t = \left(\frac{1}{N} \sum_{i=1}^{N} z_{it}^{v-1} \right)^{\frac{1}{v-1}}, N = \int_j M_j$$

$$\left(\widetilde{Z}_t \right)^{\nu-1} = \frac{1}{N} \sum_{i=1}^{N} z_{it}^{v-1}$$

$$e^{(\nu-1)\log \widetilde{Z}_t} = \frac{1}{N} \sum_{i=1}^{N} e^{(v-1)\log z_{it}}$$

Expand the LHS around $\overline{\log z_{it}} := \frac{1}{N} \sum_{i=1}^{N} \log z_{it}$:

$$\begin{aligned} e^{(\nu-1)\log \widetilde{Z}_t} &\approx e^{(\nu-1)\overline{\log z_{it}}} + (\nu-1) e^{(\nu-1)\overline{\log z_{it}}} \left(\log \widetilde{Z}_t - \overline{\log z_{it}} \right) \\ &= e^{(\nu-1)\overline{\log z_{it}}} \left[1 + (\nu-1) \left(\log \widetilde{Z}_t - \overline{\log z_{it}} \right) \right]. \end{aligned}$$

Similarly expand RHS terms around $\overline{\log z_{it}}$:

$$\begin{aligned} \frac{1}{N} \sum_{i=1}^{N_j} \left(e^{(\eta-1)\log z_{it}} \right) &= \frac{1}{N} \sum_{i=1}^{N} \left(e^{(\nu-1)\overline{\log z_{it}}} \right) + \frac{1}{N} (\nu-1) e^{(\nu-1)\overline{\log z_{it}}} \sum_{i=1}^{N} \left(\log z_{it} - \overline{\log z_{it}} \right) \\ &\quad + \frac{1}{2} \frac{1}{N} (\nu-1)^2 e^{(\nu-1)\overline{\log z_{ij}}} \sum_{i=1}^{N} \left(\log z_{it} - \overline{\log z_{it}} \right)^2 \\ &= e^{(\nu-1)\overline{\log z_{it}}} \left[1 + \frac{1}{2} (\nu-1)^2 \frac{\sum_{i=1}^{N} \left(\log z_{it} - \overline{\log z_{it}} \right)^2}{N} \right] \end{aligned}$$

Equating the LHS and RHS, we get:

$$\begin{aligned} e^{(\nu-1)\overline{\log z_{it}}} \left[1 + (\nu-1) \left(\log \widetilde{Z}_t - \overline{\log z_{it}} \right) \right] &= e^{(\nu-1)\overline{\log z_{it}}} \left[1 + \frac{1}{2} (\nu-1)^2 \frac{\sum_{i=1}^{N} \left(\log z_{it} - \overline{\log z_{it}} \right)^2}{N} \right] \\ \log \widetilde{Z}_t - \overline{\log z_{it}} &= \frac{1}{2} (\nu-1)^2 \frac{\sum_{i=1}^{N} \left(\log z_{it} - \overline{\log z_{it}} \right)}{N} \\ \log \widetilde{Z}_t &= \mathbb{E}\left[\log z_{it} \right] + \frac{1}{2} (\nu-1) \frac{\sum_{i=1}^{N} \left(\log z_{it} - \overline{\log z_{it}} \right)^2}{N} \\ \log \widetilde{Z}_t &= \mathbb{E}\left[\log z_{it} \right] + \frac{1}{2} (\nu-1) \mathbb{V}\left[\log z_{it} \right] \end{aligned} \tag{B15}$$

B.5 Markup Decomposition

This section documents the method we use to decompose markups. Following Haltiwanger (1997), we decompose the change in the markup into a component that is due to (i) changes in market shares (Δ

Reallocation), (ii) changes in markups themselves (Δ Within), and (iii) the effect of Net entry as follows:

$$\Delta\mu_t = \underbrace{\underbrace{\sum_{i,j}\widetilde{\mu}_{ij,t-1}\Delta m_{ij,t}}_{\Delta \text{ Market share}} + \underbrace{\sum_{i,j}\Delta\mu_{ij,t}\Delta m_{ij,t}}_{\Delta \text{ Cross term}}}_{\text{(i) } \Delta \text{ Reallocation}} + \underbrace{\sum_{i,j} m_{ij,t-1}\Delta\mu_{ij,t}}_{\text{(ii) } \Delta \text{ Within}} + \underbrace{\sum_{i,j\in\text{Entry}}\widetilde{\mu}_{ij,t}m_{ij,t} - \sum_{i,j\in\text{Exit}}\widetilde{\mu}_{ij,t-1}m_{ij,t-1}}_{\text{(iii) Net entry}}, \tag{B16}$$

where $\widetilde{\mu}_{ij,t} = \mu_{ij,t} - \mu_{t-1}$, and $\widetilde{\mu}_{ij,t-1} = \mu_{ij,t-1} - \mu_{t-1}$ are deviations from the economy wide markup and m_{ij} denotes the revenue share of any firm i in the entire economy.

Since the estimated parameters of our economy are different in different years, but we assume a steady state of the economy in each year, then we need to make some assumptions in order to proceed with this decomposition. The assumption we make is to map firms together over time according to their subscript ij. Let $\overline{M} = \max\{M_t\}$ be the largest number of potential entrants and let $\mathbf{U} = \{u_{ijt}\}$ be an array of uniform random numbers for all firms $i \in \{1, \ldots, \overline{M}\}$ in all j sectors in all j periods. We start in 1980. Given the stationary distribution of productivity implied by productivity process and parameters $\rho, \sigma^{\varepsilon}_{1980}$, the first set of random numbers determine initial productivity by inverting the CDF of this distribution. We then use the remaining random numbers and the sequence of productivity process parameters σ_t for 1981 to 2016 to evolve productivity forward for all $\overline{M} \times J$ firms forward at random. This gives us an array $\mathbf{Z}^* = \{z^*_{ijt}\}$ of *latent* productivities of the $\overline{M} \times J$ over the T periods.

To take care of changes in the number of potential entrants we then proceed as follows. In 1980 there are $M_{1980} \leq \overline{M}$ potential entrants. In each market, we draw at random the M_{1980} potential entrants from the $\overline{M}$ firms, and in 1980 set the remaining firms' productivity to zero. In 1981, if $M_{1981} < M_{1980}$ then we randomly select $M_{1980} - M_{1981}$ of the M_{1980} firms and set their actual productivities $z_{ijt} = 0$, while for the potential entrants we set $z_{ijt} = z^*_{ijt}$.

This then gives us a distribution of firm productivities $\mathbf{Z}_t = \{z_{ijt}\}$ in all periods, of which these are zero for firms that are not considered potential entrants. Given the parameter ϕ_t we can then solve for the steady state of the economy in each t, recording markups and sales for all firms. For firms that are not potential entrants, these are obviously both zero. We then solve this economy in each year record markups and sales shares and apply the above decomposition.

C A variance-covariance decomposition

The endogenous wedges that determine aggregate quantities in our economy $\{z_t, \mu_t, \omega_t\}$ depend on the joint distribution of productivity and markups in a way that is not completely transparent. As a final exercise in this section we show that, quantitatively, (i) the mapping from the joint distribution of productivity and markups to aggregate wedges can be summarized well by five moments that capture means, variances and the covariance of markups and productivity, (ii) that these depend mostly on the mean terms, (iii) account for how primitives determine these moments.

A single-nest approximation. Our approach is to take a second order approximation of each wedge around a particular moment of the joint distribution of productivities and markups. While this is possible in the nested case, it leads to complicated formulas that contain within and between market variance and covariance terms. For the purposes of this exercise it turns out that an excellent approximation can be obtained by aggregating markups and productivities *as if* the economy had a single nest, with elasticity of substitution ν. In this case we denote aggregates with a tilde:

$$\widetilde{Z}_t = \left[\frac{1}{\int M_j dj}\int \sum_{i=1}^{M_{jt}} z_{ijt}^{\nu-1} dj\right]^{\frac{1}{\nu-1}}, \ \widetilde{\mu}_t^{-1} = \left[\frac{1}{\int M_j dj}\int \sum_{i=1}^{M_{jt}} \left(\frac{z_{ijt}}{\widetilde{Z}_t}\right)^{\nu-1} (\mu_{ijt}^{-1})^{\nu-1} dj\right]^{\frac{1}{\nu-1}}, \ \widetilde{\Omega}_t = \int \sum_{i=1}^{M_{jt}} \left(\frac{z_{ijt}}{\widetilde{Z}_t^{\nu}}\right)^{\nu-1} \left(\frac{\mu_{ijt}}{\widetilde{\mu}_t}\right)^{-\nu} di. \tag{C1}$$

We can choose ν in different ways, all of which deliver good approximations. We choose an approach that does not use information on markups. We set ν such that $\widetilde{Z}_{2016}^{\nu} = Z_{2016}$, which delivers a value of $\nu = 4.40$. As expected this is between $\theta = 1.20$ and $\eta = 5.75$.

This approximation works well, closely matching the path for each wedge. Again using lower case to denote objects in log changes from 1980, Figure C1 plots the baseline wedges from the model in red, and the approximation using $\nu = 4.40$ in blue dashed lines.

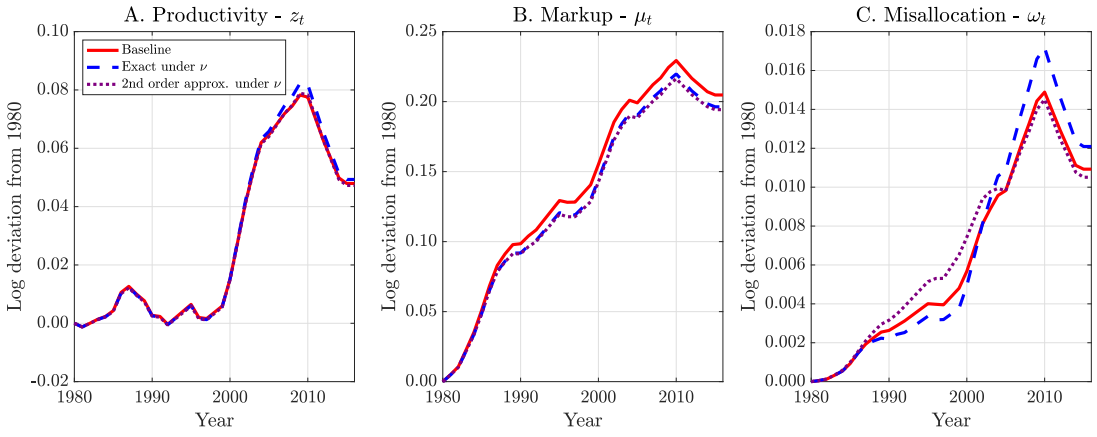

Figure C1: Single-nest and second-order approximations of z_t, μ_t and ω_t

Notes: This figure takes the baseline wedges from the model (red, solid) for productivity, the markup and misallocation, and compares them to (i) the single-nest approximating model under $\nu = 4.40$ (blue, dashed), and (ii) the second order approximation of the single-nest approximating model (purple, dotted).

Second-order approximation. The single-nest expressions (C1) admit clean second order approximations. We approximate log productivity terms around $\mathbb{E}\left[\log z_{ijt}\right]$ and log markup terms around $\mathbb{E}\left[\log \mu_{ijt}\right]$. Doing so, to a second order, we obtain the following for the wedges in log changes, $\widetilde{z}_t$ and $\widetilde{\mu}_t$:

$$\widetilde{z}_t \approx \Delta\mathbb{E}_t\left[\log z_{ijt}\right] + \frac{1}{2}(\nu-1)\,\Delta\mathbb{V}_t\left[\log z_{ijt}\right] \tag{C2}$$

$$\widetilde{\mu}_t \approx \Delta\mathbb{E}_t\left[\log \mu_{ijt}\right] - \frac{1}{2}(\nu-1)\,\Delta\mathbb{V}_t\left[\log \mu_{ijt}\right] + (\nu-1)\,\Delta\mathbb{C}_t\left[\log z_{ijt}, \log \mu_{ijt}\right]. \tag{C3}$$

While the aggregate productivity term is increasing in the variance of productivity, the aggregate markup term is decreasing in the variance of markups. This can be understood as follows. Recall that in an efficient allocation markups ($\mu_{ijt} = 1$) variable factor productivity is $\widetilde{Z}_t$. An increase in productivity dispersion reallocates factors to higher productivity firms, increasing aggregate productivity $\widetilde{Z}_t$. The higher is ν the more aggressively these factors are reallocated, boosting aggregate productivity. Meanwhile, the covariance between markups and productivity increases the aggregate markup wedge, as the contraction in factor demand relative to the efficient benchmark is more severe when the higher markups belong to higher productivity firms. The model we study generates a positive covariance through Cournot competition.

These can then be used to simplify the second order expansion of the misallocation term:

$$\widetilde{\omega}_t \approx \frac{1}{2}(\nu-1)^2\,\mathbb{V}_t\left[\log z_{ijt}\right] + \frac{1}{2}\nu^2\mathbb{V}_t\left[\log \mu_{ijt}\right] - \nu(\nu-1)\,\mathbb{C}_t\left[\log z_{ijt}, \log \mu_{ijt}\right] - (\nu-1)\left\langle \widetilde{z}_t - \Delta\mathbb{E}_t\left[\log z_{ijt}\right]\right\rangle + \nu\left\langle \widetilde{\mu}_t - \Delta\mathbb{E}_t\left[\log \mu_{ijt}\right]\right\rangle$$

This full expression shows how misallocation depends on the variance and covariance of markups and productivity. However after substituting in the above expressions for $\widetilde{z}_t$ and $\widetilde{\mu}_t$ into the terms in $\langle\cdot\rangle$, we find that $\widetilde{\omega}_t$ depends only on the variance of markups: $\widetilde{\omega}_t \approx \frac{\nu}{2}\Delta\mathbb{V}_t\left[\log \mu_{ijt}\right]$.

The purple dotted line in Figure C1 show that the second order approximation does well in capturing the aggregate wedges, in particular productivity and markup, which we have shown are quantitatively the important wedges for understanding aggregate moments (Figure 10). To complete the picture, Appendix Figure C4 combines the decomposition of output and welfare into wedges, with the decomposition of the wedges into moments of the joint distribution of log productivity and log markups. As expected, the mean terms dominate in terms of markups, with the increasing covariance contributing only one tenth of the decline in output and welfare.

Results. Figure C2 plots the contribution of each of the five moments to the wedges in equations (C2), (C3) and (C4). Our main result is that while the level and variance of productivity are equally important for understanding the aggregate productivity wedge, the markup can be well summarized by the mean of markups while, by our above result, the misallocation term depends only on the variance of markups. Quantitatively, the key result is that the covariance of markups and productivity, which increases over time, contributes less than a quarter of the increase in the markup wedge, and is partially offset by the increase in the variance.

Figure C3 shows how each parameter contributes to the moments in Figure C2. First, and consistent

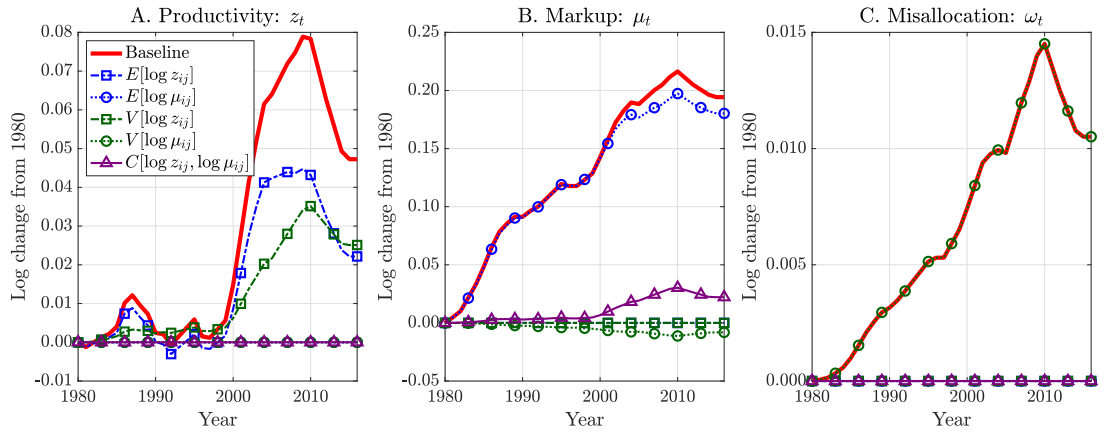

Figure C2: Decomposing wedges into moments of the joint distribution of markups and productivity

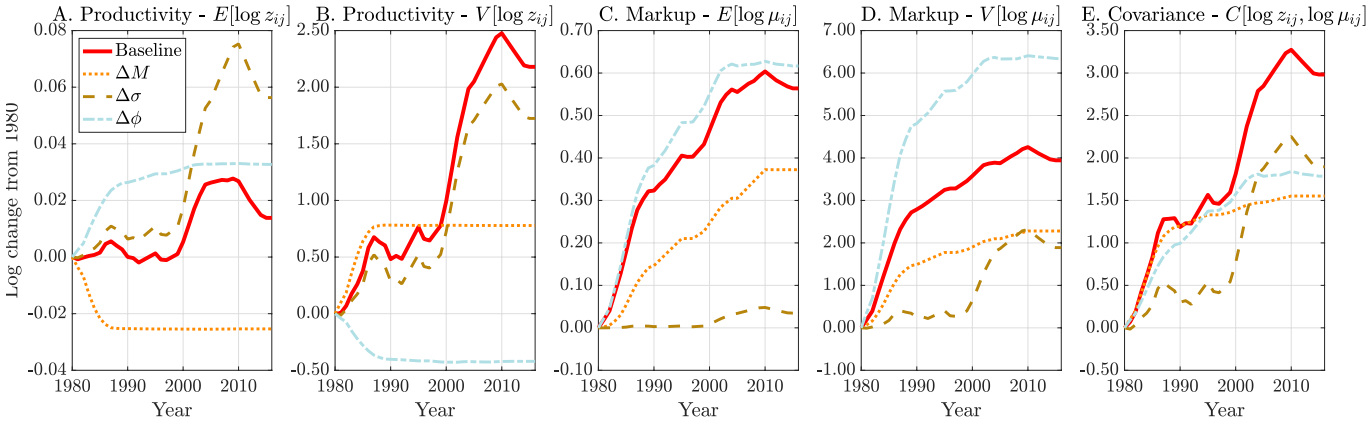

Figure C3: Effect of each parameter on the joint distribution of markups and productivity

Notes: This figure plots the effects of changing parameters independently on each of the five moments plotted in Figure C2. For example, the orange dotted lines with square markers give the time-series of wedges implied by feeding only the estimated time-series for $\{M_t\}_{t=1980}^{2016}$ into the model, keeping ϕ_t and σ_t fixed at their estimated 1980 values.

with our results from the previous section, the off-setting forces of the decline in competition and increase in fixed costs leave the dispersion in productivity to shape the mean and variance of productivity. Second, the time-series for the arithmetic mean log markup, which shapes $\widetilde{\mu}_t$, is determined by both market structure and technology, with the decline in M_t leading to a dampening of the selection effects that would occur under only an increase in ϕ_t. Although quantitatively not important for the aggregate wedges that determine output, employment and welfare, we note that the covariance of markups and productivity is shaped by all three parameters. All three forces increase the market power of the most productive firms in the economy, increasing this covariance.

Fixed cost. The aggregate overhead wedge $\widetilde{\Phi}_t$ that enters aggregate TFP also depends on the distribution of firms as well as the time-varying estimate of ϕ_t. The fixed cost wedge Φ_t can be written

$$\widetilde{\Phi}_t := \left(\frac{N_t}{N_t - \Phi_t}\right)^{\alpha} = \left[\int \sum_{i=1}^{M_{jt}} \underbrace{\left(\frac{n_{it,Prod}}{N_{t,Prod}}\right)}_{\text{Weight: } \xi_{it}} \phi_{it}^{\alpha}\right]^{1/\alpha} \quad , \quad \widetilde{\phi}_{it} = \left(\frac{n_{it}}{n_{it} - \phi_t}\right)^{\alpha}. \tag{C4}$$

A second order approximation delivers $\widetilde{\phi}_t = \Delta \log \widetilde{\Phi}_t$, where

$$\widetilde{\phi}_t \approx \Delta \mathbb{E}\left[\log \widetilde{\phi}_{it}\right] + \frac{1}{2\alpha}\Delta \mathbb{V}\left[\log \widetilde{\phi}_{it}\right] + \frac{\alpha}{2}\Delta \mathbb{V}\left[\log \xi_{it}\right] + \Delta \mathbb{C}\left[\log \widetilde{\phi}_{it}, \log \xi_{it}\right]. \tag{C5}$$

Figure C5A shows that as with the previous approximations, this closely matches the true time-series for the wedge. Figure C5B shows that, this is driven almost entirely by the change in the mean of firm level $\log \widetilde{\phi}_{it}$. Consistent with our previous results, Figure C6 shows that if only M were to decline, then this term would have declined, while in the presence of an increase in ϕ, the net effect is positive.

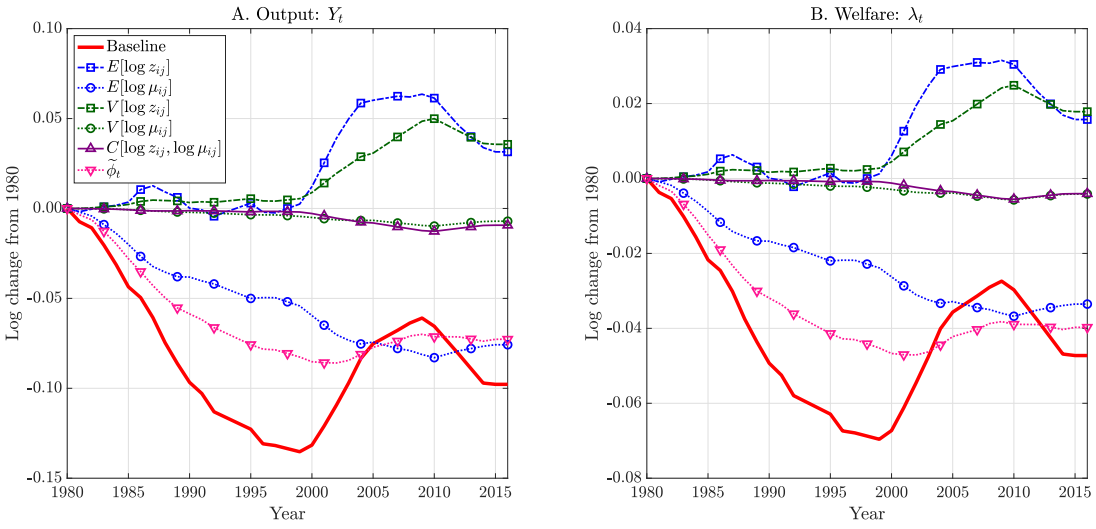

Figure C4: Decomposing output and welfare into variance and covariance of markups and productivity

Notes: This figure combines the second order approximation that decomposes wedges into moments of the joint distribution of log productivity and log markups (equations (C2) to (C4)) and the expressions that decompose output (28) and welfare (29) into wedges, to decompose output and welfare into the moments of the joint distribution of log productivity and log markups.

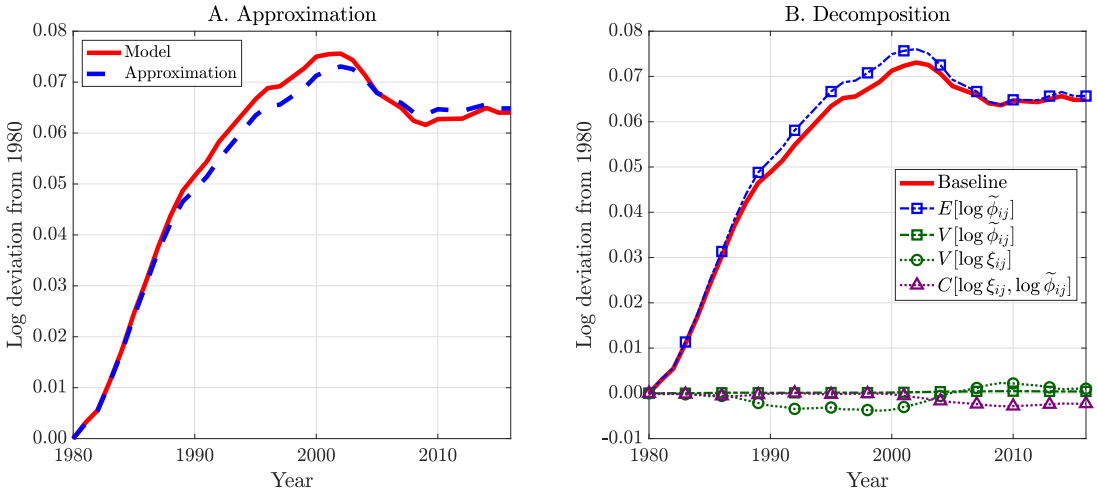

Figure C5: Approximation and decomposition of overhead wedge $\widetilde{\Phi}_t$

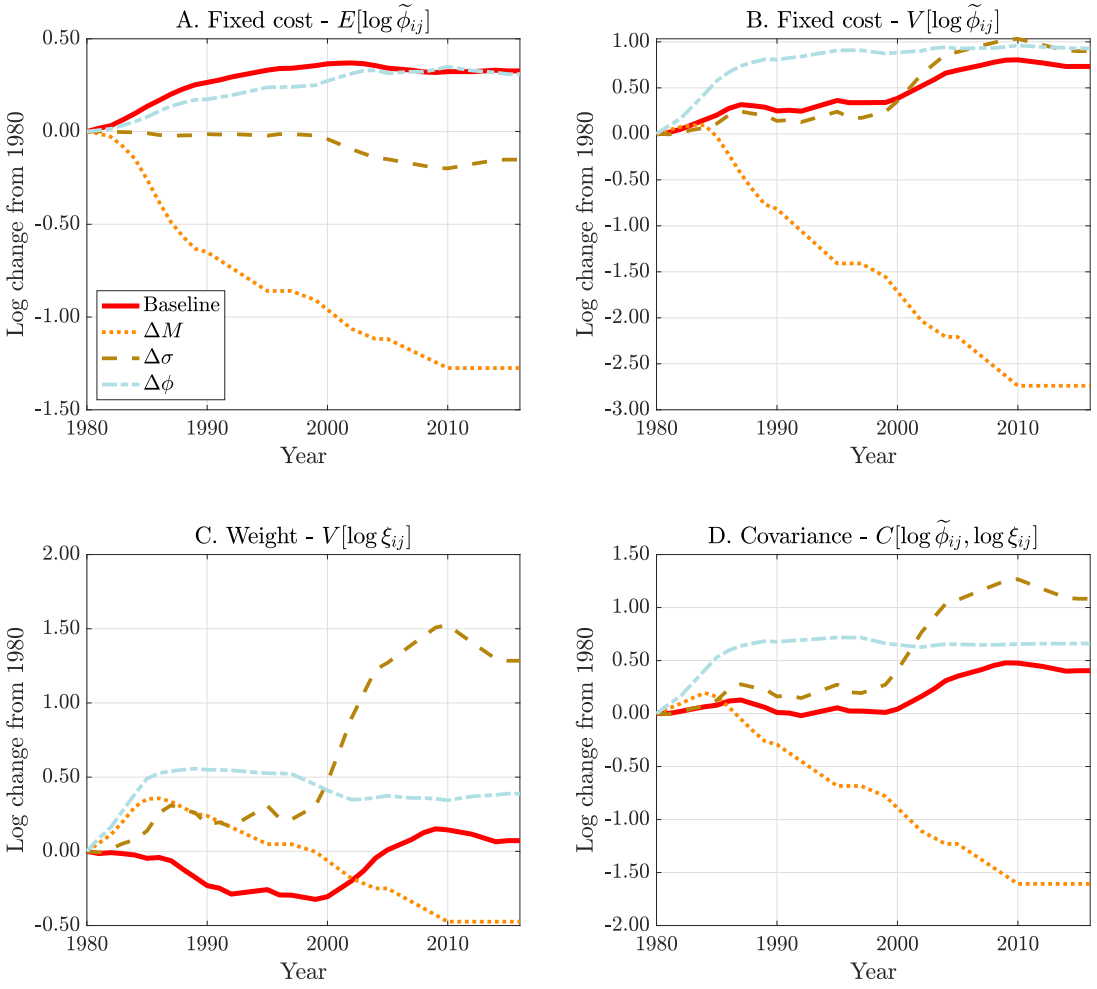

Figure C6: The effects of parameters on statistical moments of the decomposition of overhead

D Additional figures and tables

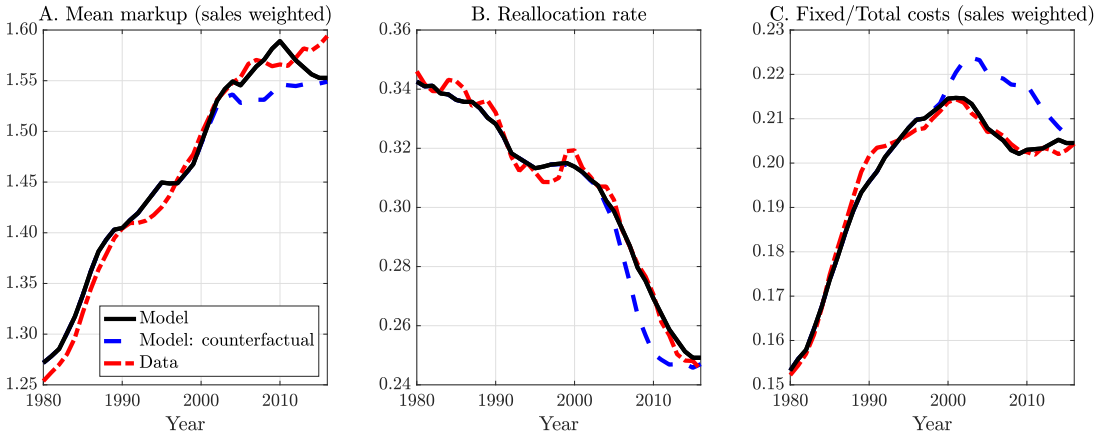

Figure D1: Model fit - Moments with a counterfactual path for σ_t

Notes: This replicates Figure 3, with the addition of the blue dashed line which corresponds to moments under the counterfactual path for σ in Figure D2, below.

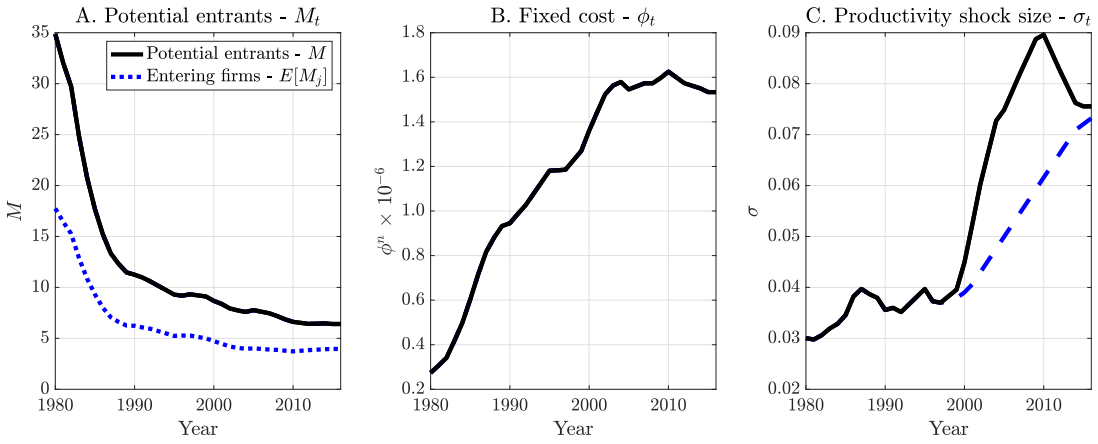

Figure D2: Parameter estimates with a counterfactual path for σ_t

Notes: This replicates Figure 4, with the addition of the blue dashed line in panel C which corresponds to moments a counterfactual path for σ which smoothly joins the estimated path for σ between its 1998 and 2016 values.

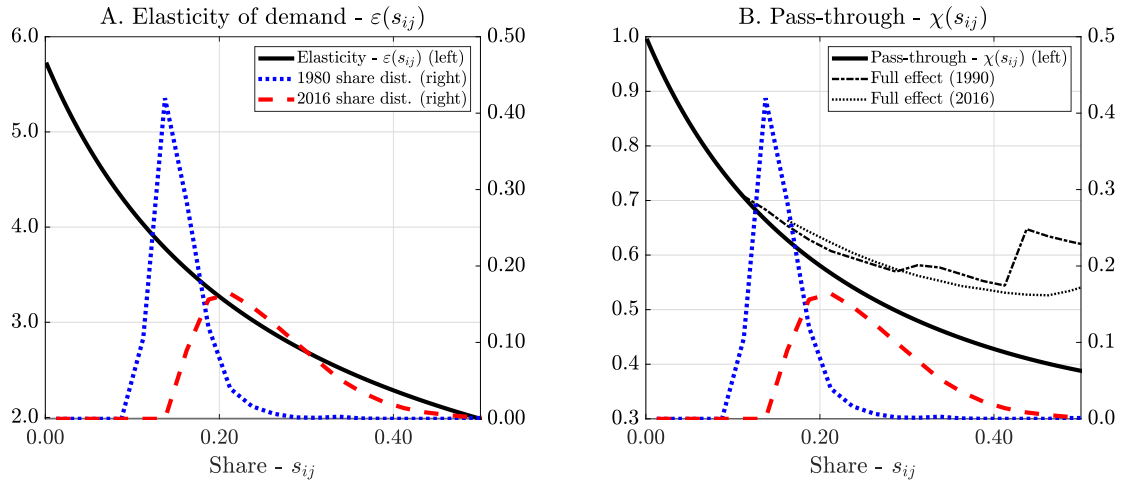

Figure D3: Elasticity of demand and pass-through by share in the model, 1980 and 2016

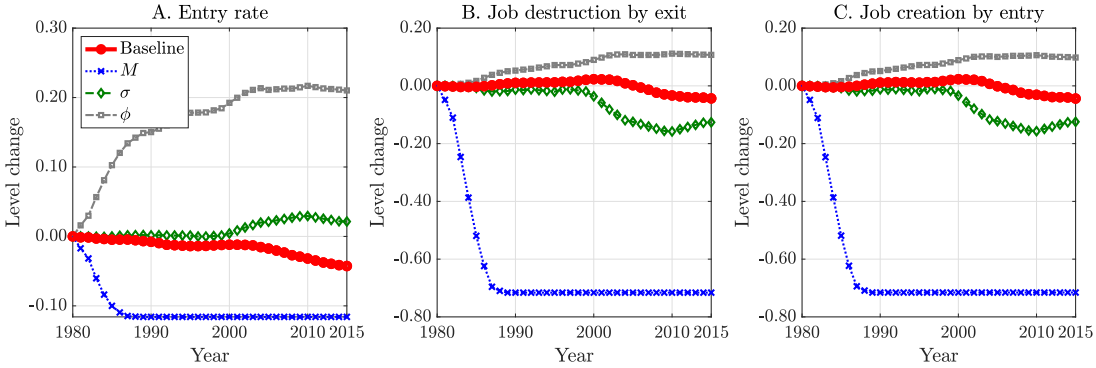

Figure D4: The effects of market structure and technology on entry rate and labor dynamism

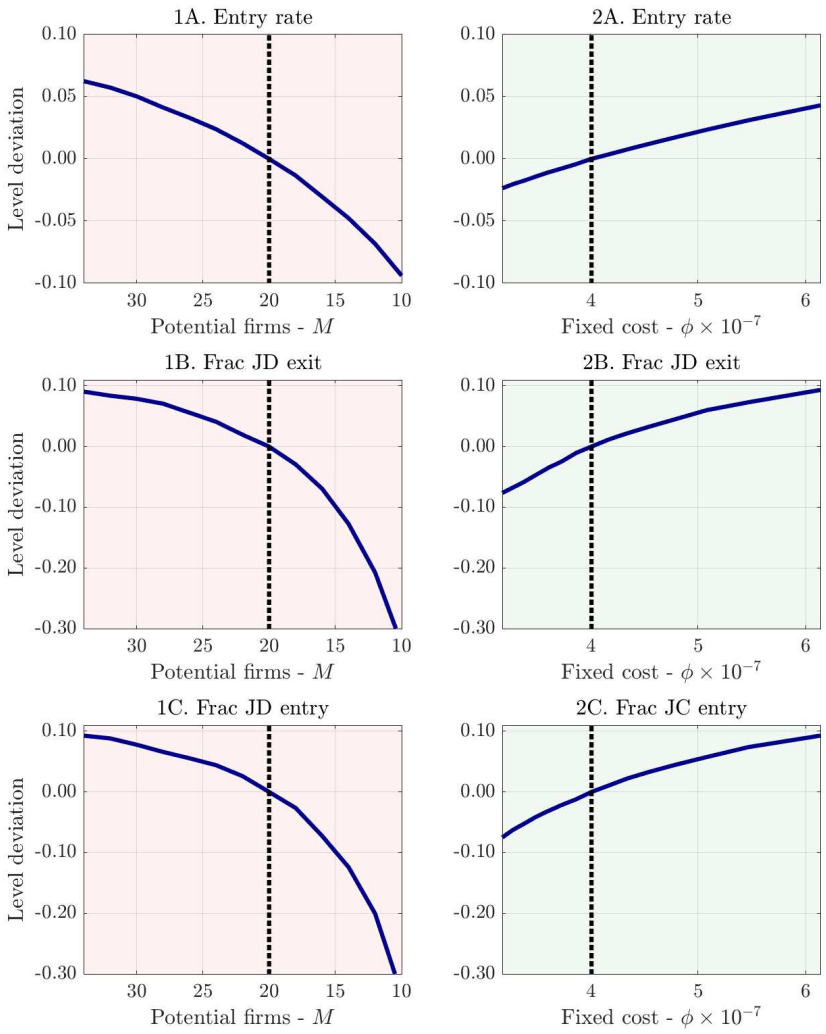

Figure D5: Entry rate and the composition of job destruction due to M and ϕ

Notes: The format of this plot follows from Figure 2. The vertical axis plots the variables of interest in level deviations from their value when each parameter is set to its median value between 1980 and 2016, which is marked by the black dashed line. Each parameter is then varied between its minimum and maximum value over 1980 to 2016.

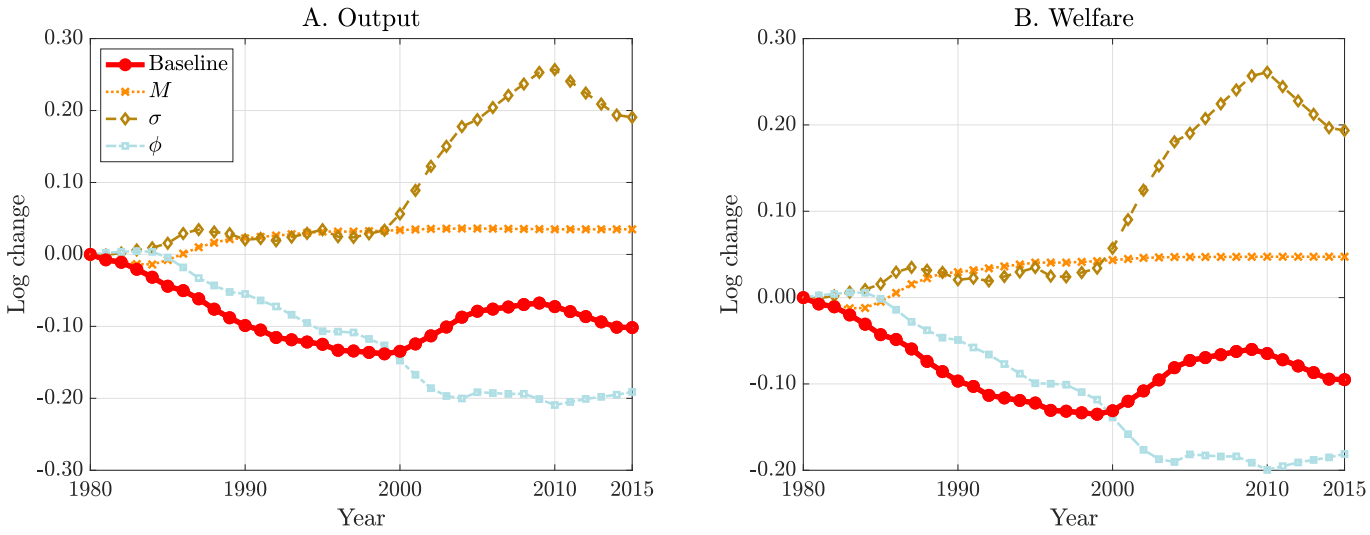

Figure D6: Effect of each parameter on output and welfare

Notes: dotted lines v into the mc

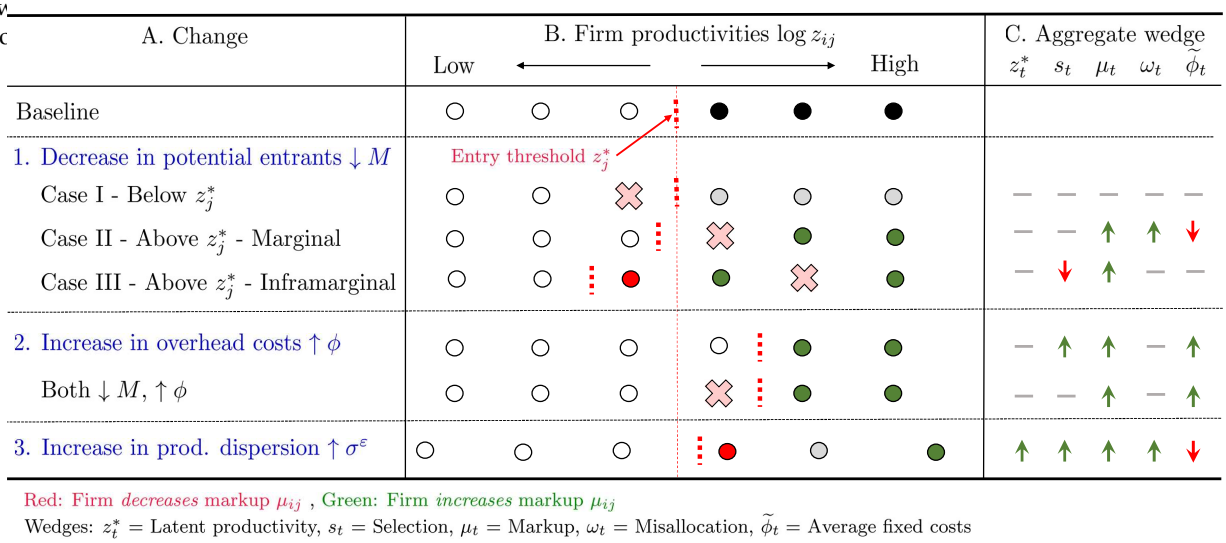

Figure D7: Parameter changes and wedges

Notes: This diagram shows how changes in each parameter, and combinations of parameters affect the 5 wedges that we have shown are determinative of aggregates in the economy. A baseline sector with 6 firms is given in the first row, with the red vertical dashed line denoting the market-specific, endogenous productivity threshold for entry. In Panel 1, we consider three possible cases of a decrease in the number of potential firms. As an example, in the final row, the dispersion of productivity increases. Productivity wedge z^* increases due to convexity, selection s^* increases as incumbent productivity increases relative to non-incumbents, markup wedge μ increase due to more market power at the highest productivity firm, misallocation wedge ω increases as the highest productivity firms' markup increases, while fixed cost wedge $\widetilde{\phi}$ declines as firms expand on the intensive margin.

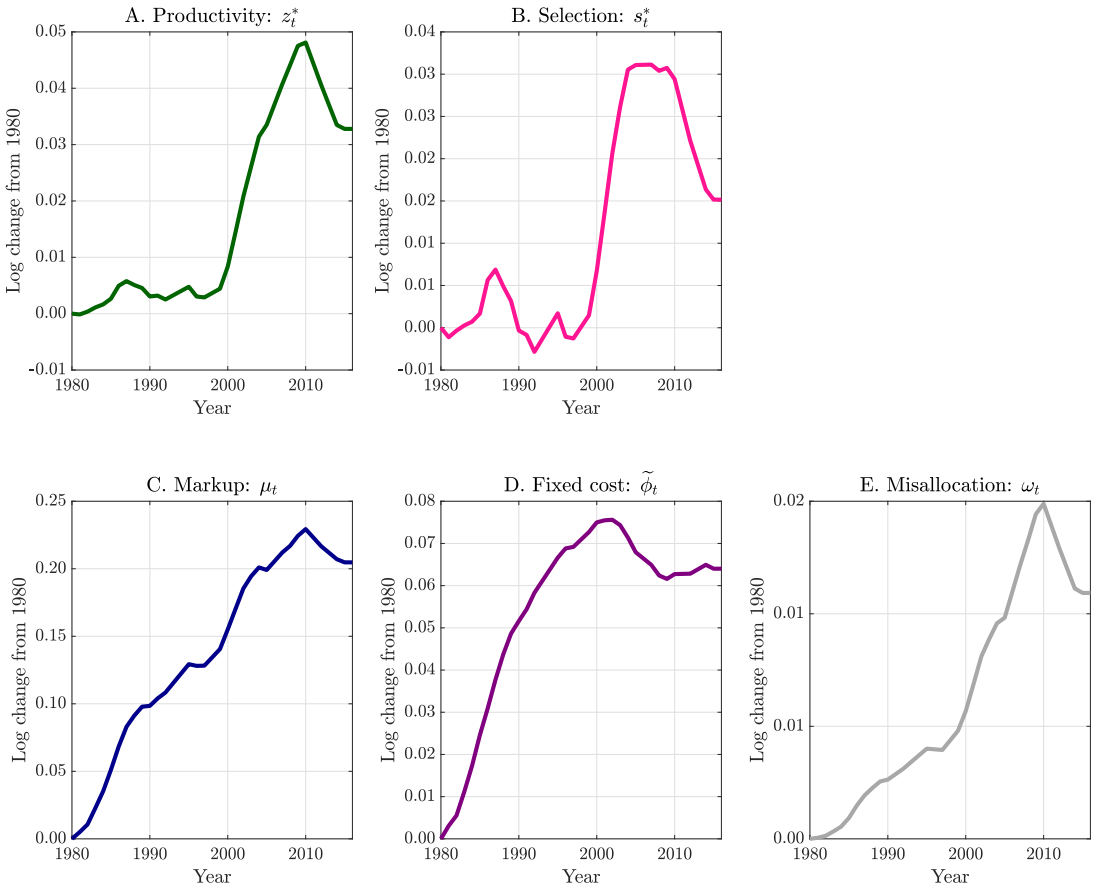

Figure D8: Wedges - time series

Notes: This figure plots the time-series of the wedges that appear in the set of general equilibrium conditions of the model.

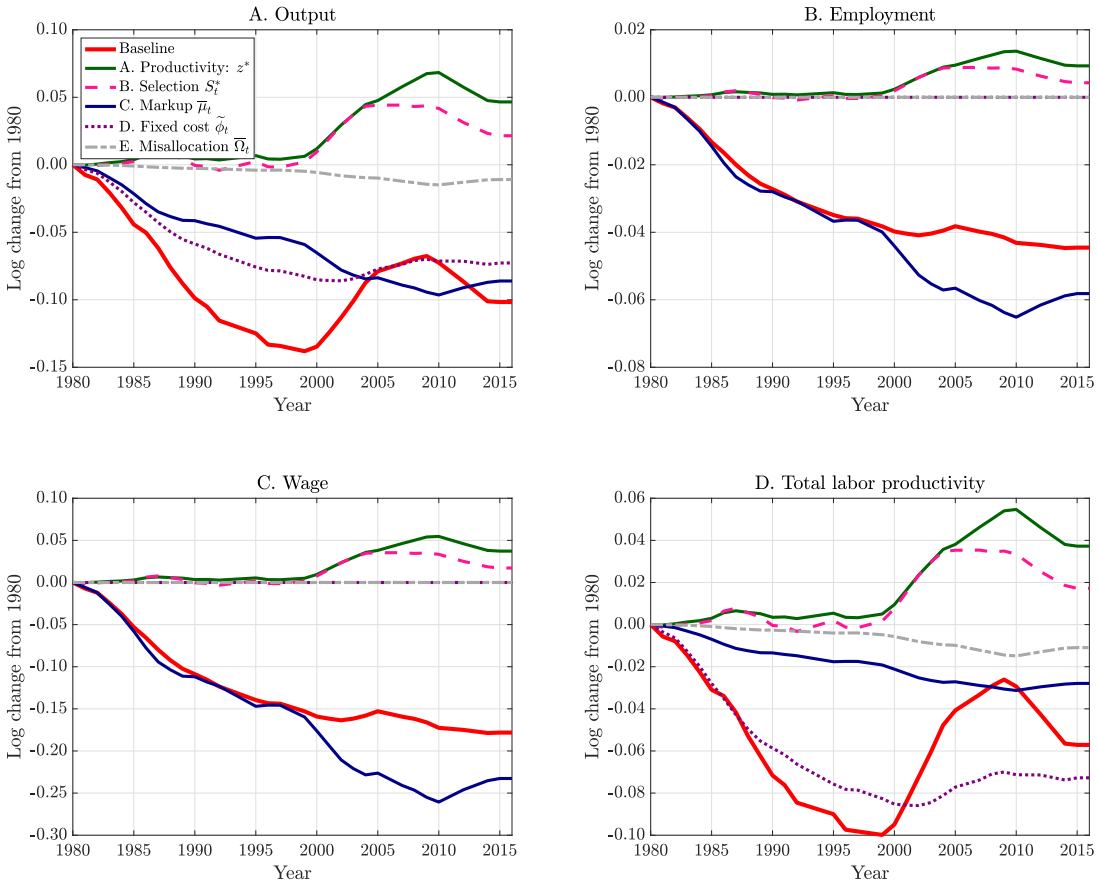

Figure D9: The effect of aggregate equilibrium wedges on aggregate quantities and prices

Notes: For context, Panel A replicates Figure 10 from the main text. Panels B, C, and D repeat the same exercise for employment n_t, the wage w_t and total labor productivity which is defined as $y_t - n_t$

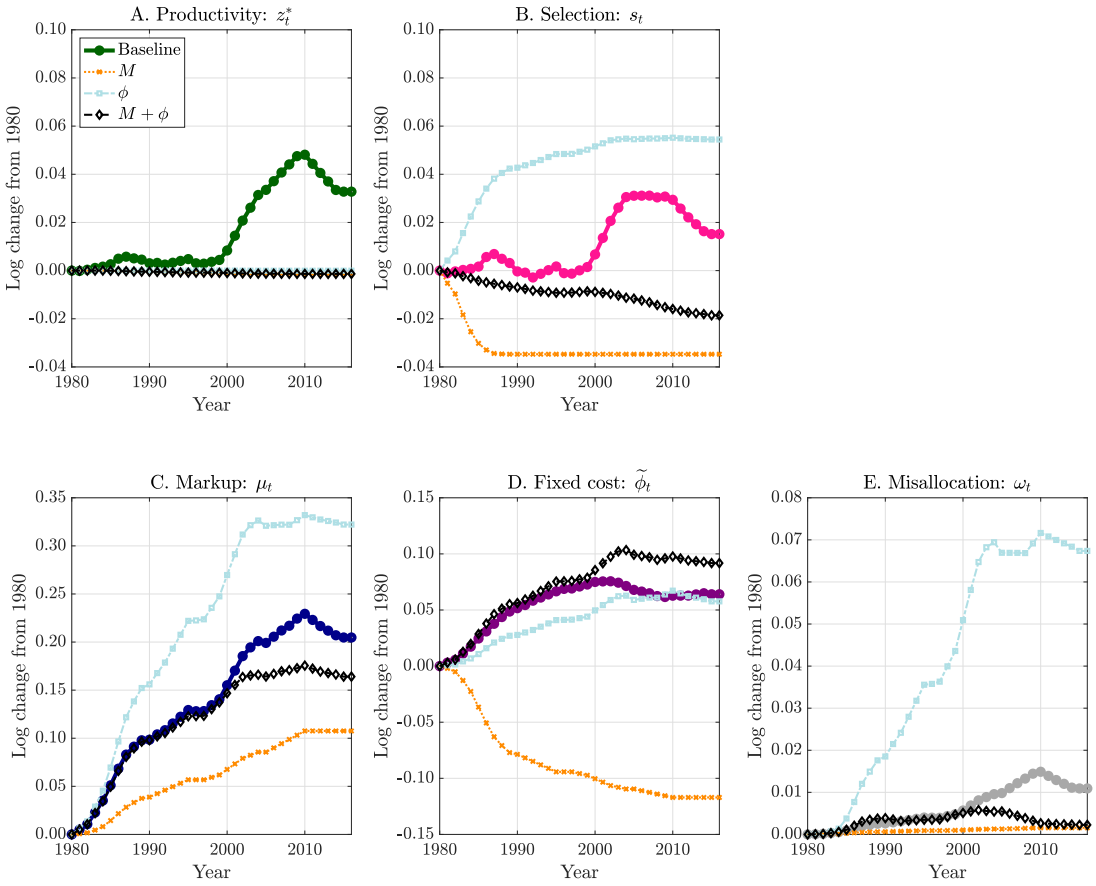

Figure D10: Effect of M, ϕ and the two combined (black diamonds)

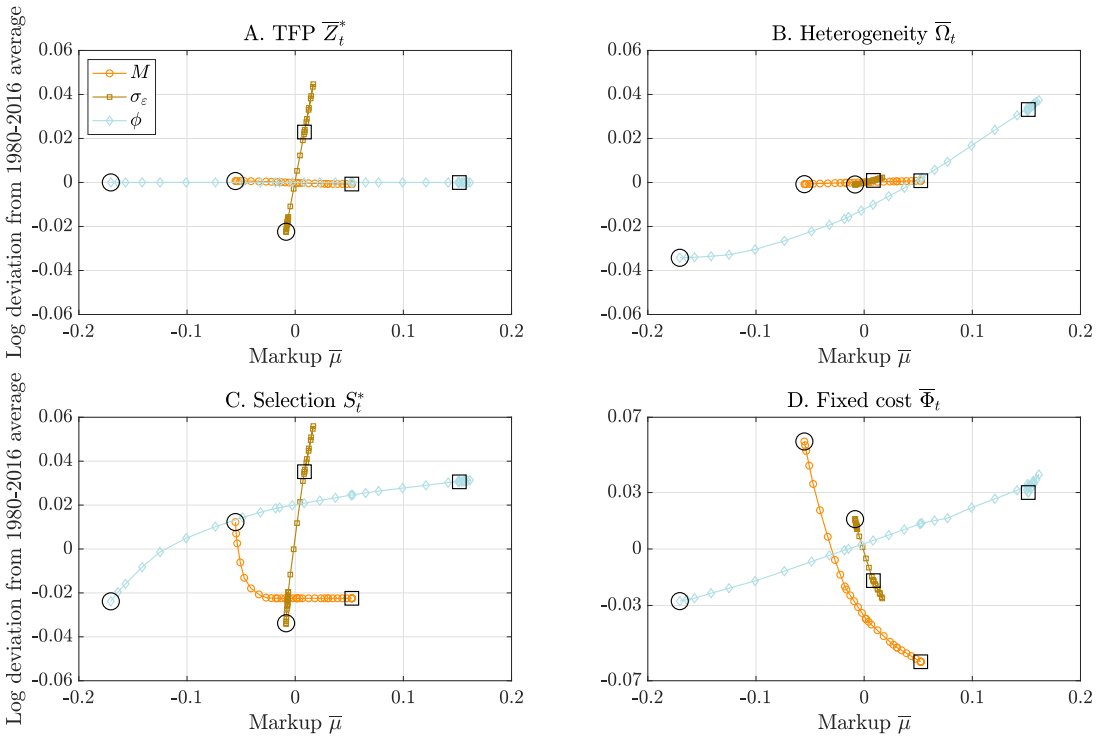

Figure D11: Covariance structure of wedges due to different parameters

Notes: This figure plots the time-series of the wedges $\{z_t^*, s_t, \omega_t, \phi_t\}$ from the model against the time-series of the markup wedge μ_t, that are induced by changes in each parameter separately. Each series is plotted in log deviations from the average over 1980 to 2016. For cross-reference, the colors here match Figure 12. The circle marker corresponds to the 1980 values of the parameters, while the square corresponds to the 2016 value.