THE
QUARTERLY JOURNAL
OF ECONOMICS

Vol. 135 2020 Issue 2

THE RISE OF MARKET POWER AND THE MACROECONOMIC IMPLICATIONS*

JAN DE LOECKER
JAN EECKHOUT
GABRIEL UNGER

We document the evolution of market power based on firm-level data for the U.S. economy since 1955. We measure both markups and profitability. In 1980, aggregate markups start to rise from 21% above marginal cost to 61% now. The increase is driven mainly by the upper tail of the markup distribution: the upper percentiles have increased sharply. Quite strikingly, the median is unchanged. In addition to the fattening upper tail of the markup distribution, there is reallocation of market share from low- to high-markup firms. This rise occurs mostly within industry. We also find an increase in the average profit rate from 1% to 8%. Although there is also an increase in overhead costs, the markup increase is in excess of overhead. We discuss the macroeconomic implications of an increase in

*We thank Mark Aguiar, Pol Antràs, John Asker, Eric Bartelsman, Susanto Basu, Steve Berry, Tim Bresnahan, Emmanuel Farhi, Xavier Gabaix, Bob Hall, John Haltiwanger, Eric Hurst, Loukas Karabarbounis, Patrick Kehoe, Pete Klenow, Christian Michel, Ariel Pakes, Thomas Philippon, Esteban Rossi-Hansberg, Chad Syverson, James Traina, Jo Van Biesebroeck, and Frank Verboven for insightful discussions and comments. Shubhdeep Deb provided invaluable research assistance. De Loecker gratefully acknowledges support from the FWO Odysseus Grant, and the ERC, Consolidator Grant 816638, and Eeckhout from the ERC, Advanced Grant 339186, and from ECO2015-67655-P. This article uses restricted data that were analyzed at the U.S. Census Bureau Research Data Center in Boston. Any opinions and conclusions expressed herein are those of the authors and do not necessarily represent the views of the U.S. Census Bureau. All results have been reviewed to ensure that no confidential information is disclosed. The authors declare that they have no relevant or material financial interests that relate to the research described in this article.

The Quarterly Journal of Economics (2020), 561–644. doi:10.1093/qje/qjz041.
Advance Access publication on January 23, 2020.

average market power, which can account for a number of secular trends in the past four decades, most notably the declining labor and capital shares as well as the decrease in labor market dynamism. *JEL* Codes: E2, D2, D4, J3, K2, L1.

I. Introduction

Thriving competition between firms is a central tenet of a well-functioning economy. The pressure of competitors and new entrants leads firms to set prices that reflect costs, which is to the benefit of the customer. In the absence of competition, firms gain market power and command high prices. This has implications for welfare and resource allocation. In addition to lowering consumer well-being, market power decreases the demand for labor and dampens investment in capital, it distorts the distribution of economic rents, and it discourages business dynamics and innovation. This has ramifications for policy, from antitrust to monetary policy and income redistribution.

Despite the vital importance of market power in economics, surprisingly little is known about its systematic patterns for the aggregate economy and over time. In this article, our main goal is to document the evolution of market power for the U.S. economy since the 1950s. First, we analyze markups, the most common measure of whether firms are able to price their goods above marginal cost. Traditionally in the industrial organization literature, this measure is of importance because it is informative about the technology that firms use and whether there is efficiency in production. Based on firm-level data, we find that although aggregate markups were more or less stable between 1955 and 1980, there has been a steady rise since 1980, from 21% above cost to 61% above cost in 2016. More important than the increase in the aggregate markup, the main insight is that the distribution of markups has changed: the median is constant, and the upper percentiles have gone up substantially. This rise in markups by a few firms has gone together with reallocation of economic activity. A few firms have high markups and are large, but the majority of firms see no increase in markups and lose market share.

Markups alone do not tell the full story about market power. For example, markups may be high because overhead costs or fixed costs are high. In that case, the firm charges prices well above marginal costs to cover fixed costs. We therefore also analyze measures of profitability that take into account not only the marginal cost but total costs, including the expenditure on capital

and fixed or overhead costs. Because measuring profitability is challenging, we document a rise in different measures, ranging from accounting profits to stock market performance. We show that both measures, markups and profitability, are related. Although we do find that there is an increase in overhead costs, the rise of markups cannot exclusively be attributed to overhead. Markups have gone up more, and as a result, so has profitability. The increase in both markups and profitability provides evidence that market power has increased.

Once we have robustly established the facts, we discuss the macroeconomic implications of this rise in market power and the general equilibrium effects it has. We argue that the rise in market power is consistent with several secular trends in the past four decades, most notably the decline in the labor and capital shares, as well as the decrease in business dynamism and labor reallocation.

Measuring market power is notoriously hard. The most widely used measures of market power such as concentration ratios, for example, the Herfindahl-Hirschman Index (HHI), have serious pitfalls because they are sensitive to the definition of a market. This is especially problematic when analyzing market power in the aggregate across different industries and over long time periods, where market definitions change. Although HHI is a good measure under certain circumstances—especially when the market definition is stable and when firms compete à la Cournot for example—and is widely used, it is not an adequate measure of market power for the macroeconomy across time and space.[1]

The evidence on market power we have to date comes from case studies of specific industries,[2] for which researchers have access to detailed data. In this approach championed by Bresnahan (1989) and Berry, Levinsohn, and Pakes (1995), the estimation of markups traditionally relies on assumptions on consumer behavior coupled with profit maximization, and an imposed model of how firms compete, for example, Bertrand-Nash in prices or Cournot quantity competition. The fundamental challenge that this approach confronts is the notion that marginal costs of production are fundamentally not observed, requiring more structure to uncover them from the data. This approach requires a

1. See Bresnahan (1989) and Syverson (2019).

2. For example, cars (Berry, Levinsohn, and Pakes 1995), breakfast cereal (Nevo 2001), or beer (Koujianou Goldberg and Hellerstein 2012).

combination of data on consumer demand (containing prices, quantities, characteristics, consumer attributes, etc.) and the need for specifying a model of conduct. All these requirements have limited the use of the so-called demand approach to particular markets and prohibit its applicability for macroeconomic questions.

In this article, we follow a radically different approach to estimate markups, the so-called production approach. Building on Hall (1988), recent advances in the literature on markup estimation by De Loecker and Warzynski (2012) rely on individual firm output and input data and posit cost minimization by producers. A measure of the markup is obtained for each producer at a given point in time as the wedge between a variable input's expenditure share in revenue (directly observed in the data) and that input's output elasticity. The latter is obtained by estimating the associated production function. The advantage of this approach is twofold. First, the production approach does not require us to model demand or specify conduct for many heterogeneous markets over a long period of time. Second, we can rely on publicly available accounting data. In particular, most of the information we need is available in the financial statements of firms. Although there still exist many measurement issues and associated econometric challenges, to our knowledge there is no viable alternative to make progress on backing out economy-wide measures of market power.[3]

This article starts by documenting the main patterns of markups in the U.S. economy over the past six decades, and in doing so we provide new stylized facts on the cross-section and time-series of markups. The main analysis focuses on data from the financial statements of all publicly traded firms covering all sectors of the U.S. economy over the period 1955–2016.[4] Although publicly traded firms are relatively few compared with the total number of firms, they tend to be large. As of 2000, they account for 29% of total U.S. private sector employees, excluding the self-employed and farm workers (Davis et al. 2007). We also perform our analysis on census data where for selected industries

3. While the approaches—the demand approach and the production approach—differ, the obtained estimates should be similar. In Online Appendix 7 we compare estimates from the literature using the demand approach to our estimates for the corresponding sector.

4. The data are from Compustat, who extract the information from the Security and Exchange Commission (SEC) required public filing of financial statements. A handful of private firms are also included that have filing requirements.

we have the universe of firms.[5] We find that the distribution of markups changes dramatically since 1980: most firms see no rise in markups, whereas those in the upper tail experience a sharp rise. At the same time, there is a reallocation of economic activity toward high-markup, large firms, consistent with the superstar firm effect that Autor et al. (2020) find.

We then analyze firm profitability. The objective is to analyze whether markups have not increased exclusively because of a rise in overhead costs.[6] To address this issue, we calculate the profit rate, which is total sales minus all costs (including overhead and the expenditure on capital) as a share of sales. We find that the average profit rate has risen from close to 1% in 1980 to around 8% in 2016. While overhead costs have increased from 15% to 21% of total cost, markups have increased even more, and firms charge an excess markup that more than compensates for overhead. In fact, we find that the firms with the highest overhead costs charge the highest excess markup and therefore have the highest profits. Like markups, the increase in the average profit rate is driven by a change in the distribution, especially the upper tail. We also find that the stock market valuation as a share of sales has risen over the same period. These facts confirm that firms increasingly exert market power: they charge higher prices not just to compensate for higher overhead costs; they also obtain higher profits.

After we establish the main facts, we discuss the implications of the rise in market power for recent debates in the macro/labor literature. In particular, we analyze how the rise in markups naturally implies a decrease in the labor share. It follows immediately from the firm's optimization decision that high markups necessarily lead to lower expenditure on inputs such as labor. Hence the negative relation between markups and the labor share. We find

5. The only other attempts at measuring markups economy-wide that we have found in the literature are based on industry-level aggregate data for the period up to the 1980s. Both Burnside (1996) and Basu and Fernald (1997) find little evidence of market power (nor of returns to scale nor externalities), which is consistent with our finding that market power only picks up after 1980.

6. Although we find that the rise in markups has been accompanied by a rise in market power, even if the rise in aggregate markups we document here was purely a function of rising overhead costs and came with no change in market power, this finding would still be deeply significant. Markups are a fundamental variable throughout macroeconomics—from the benchmark New Keynesian model, to any standard endogenous growth model—as they are central to understanding technology, the efficient allocation of resources between firms, and how we think about trade-offs between static and dynamic efficiency.

that due to reallocation of economic activity toward high-markup firms, the decline in the economy-wide labor share is predominantly driven by large, high-markup firms that have individually low labor shares. This is consistent with the findings in Autor et al. (2020) and Kehrig and Vincent (2017) that large firms drive the decline in the aggregate labor share. Our finding is a slightly nuanced version: market power as a common cause determines both the increase in firm size and the decline in the labor share.

We further discuss the role of rising markups in the decrease in the capital share, the decrease in low-skilled wages, the decrease in labor market participation, and the decrease in labor reallocation and in interstate migration. The analysis of markups and market power plays a central role in many literatures in economics, most notably in industrial organization, macroeconomics, and labor economics. As a result, it has always received due attention. Currently, there are several papers that touch on the aggregate dimension of market power that we stress here. Gutiérrez and Philippon (2017) analyze the HHI of concentration as a measure of market power (see also Grullon, Larkin, and Michaely 2016 and Brennan 2016). They find that the increase in concentration is mainly driven by a decrease in domestic competition. This in turn leads to a decrease in firm-level investment, particularly in intangible assets by industry leaders. Our findings are consistent with theirs. Methodologically, our approach has the advantage that it derives firm-level markups, which circumvents the limitations of the HHI measure.[7]

Hartman-Glaser, Lustig, and Zhang (2016), Autor et al. (2020), and Kehrig and Vincent (2017) focus on the role of large firms. Hartman-Glaser, Lustig, and Zhang (2016) document that the firm-level capital share has decreased on average, even though the aggregate capital share for U.S. firms has increased. They explain the divergence with the fact that large firms now produce a larger output share even if the labor compensation has not increased proportionately. Autor et al. (2020) show the growing importance of large firms that dominate the market. They show that this leads to higher concentration and decreases the labor share, as also shown by Kehrig and Vincent (2017). Like our work,

7. Most notably, concentration is not necessarily related to market power when products are differentiated (see Bresnahan 1989), and an adequate concentration measure requires precise knowledge of what constitutes a market with information on all firms in that market.

their results are based on firm-level data, not macroeconomic aggregates.

We share with these papers that the reallocation of economic activity toward large firms has substantial implications that resolve a number of puzzles in macroeconomics, most notably the decline in the labor share. We argue that market power and the rise of markups is the common cause of both the reallocation toward large firms and the decline in the labor share. The decline in the labor share holds at the firm level, from firm optimization: as markups increase, firms spend less on labor. With an economy-wide increase in market power, enough firms reduce their expenditure on labor, which translates into an aggregate decline in the labor share, as observed in the macro aggregates.

In this article, we focus on robustly establishing the facts regarding the evolution of market power and are agnostic about the origins of the rise in market power and the corresponding reallocation of economic activity toward high-markup firms. The most prominent explanations are technological change and the change in the market structure (for example, due to the decline in antitrust enforcement, as argued by Gutiérrez and Philippon 2018). In a companion paper, De Loecker, Eeckhout, and Mongey (2018) derive quantitatively that both technological innovation and a change in market structure are at the root cause of the rise in market power. Ex ante, the effect on welfare is ambiguous: large, high-markup firms are more productive, but they extract more rents from the customer and affect the labor market adversely through lower wages. In our quantitative exercise, we find that the net effect is negative.

Finally, while we focus exclusively on the United States, there is evidence of a rise in market power around the world. Using data on publicly traded firms around the world, in De Loecker and Eeckhout (2018a) we find remarkably similar patterns of the rise in market power since 1980. As for U.S. publicly traded firms, there is a sharp rise between 1980 and 2000, a period of stagnating markups in the 2000s, followed by another sharp rise starting around 2010. The markup for the publicly traded firms increases from 1.1 in 1980 to 1.6 in 2016.

II. Empirical Framework and Data

We present the empirical framework that allows us to derive a markup measure for each firm covering the entire economy, over

more than six decades. The framework uses the cost minimization approach, where firms choose the optimal bundle of variable inputs of production. This reasonable assumption on firm behavior only relies on firm-level revenue and input expenditure data for firms across the U.S. economy. As such we do not impose restrictions on product market competition and consumer demand.

In this section, we present the model and then discuss the particular implementation in the data sets we use. Our focus is to provide a robust description and analysis of markups across producers using different methods and approaches.

II.A. Obtaining Markups from Producer Behavior

The markup is commonly defined as the output price divided by the marginal cost. Measuring markups is notoriously hard as marginal cost data is not readily available, let alone prices. There exist three distinct approaches to measure markups. First, the accounting approach relies on directly observable gross (or net) margins of profits. Although this approach is straightforward to implement, it suffers from well-known problems, chief among them the inability to directly measure the marginal cost of production. A straightforward way to circumvent this problem is to equate average to marginal costs, but this imposes strong and unrealistic restrictions on firm-level cost structures.[8]

The second approach was developed in the modern industrial organization literature (see Berry, Levinsohn, and Pakes 1995; Bresnahan 1989) and relies on the specification of a demand system that delivers price elasticities of demand. Combined with assumptions on how firms compete, the demand approach delivers measures of markups through the first-order condition associated with optimal pricing. This approach, while powerful in other settings, is not applicable here for two distinct reasons. First, we do not want to impose a specific model of how firms compete across a large data set of firms active in very different industries, or commit to a particular demand system for all the products under consideration. Second, even if we wanted to make all these assumptions, there is simply no information on prices and quantities at the product level for a large set of sectors of the economy over a long period of time. This is necessary to successfully

8. See Karabarbounis and Neiman (2018) for a recent implementation of this approach. The discussion around the merits of the use of accounting markups (and profits) dates back to Bresnahan (1989).

estimate price elasticities of demand, and specify particular models of price competition for all sectors.

Instead, we rely on a third way: the production approach. This approach is based on the insight of Hall (1988) to estimate markups from the firm's cost minimization decision. Hall (1988) used industry aggregates; De Loecker and Warzynski (2012) recently proposed to estimate firm-level markups. The method uses information from the firm's financial statements and does not require any assumptions on demand and how firms compete. Instead, markups are obtained by exploiting cost minimization of a variable input of production. This approach requires an explicit treatment of the production function to obtain the output elasticity of at least one variable input of production.

Before we discuss the production approach, on which we rely to measure markups, it is instructive to go back to the underlying assumptions of the accounting and so-called demand approaches. Throughout we define markups as the price-to-marginal cost ratio:

$$\mu \equiv \frac{P}{c}. \tag{1}$$

In essence, the simplicity of the accounting approach is to simply multiply through by total output (Q) and obtain:

$$\frac{P}{c} = \frac{PQ}{cQ}. \tag{2}$$

The entire approach rests on the assumption that the object cQ is directly observable in the data. There are three main assumptions and therefore complications. First, this approach relies crucially on the equality of marginal and average cost of production. This requires constant returns to scale (CRS) in production and the absence of economies of scale, that is, there are no fixed costs. Second, it implicitly relies on the assumption that all relevant factors of production are perfect substitutes in production. Third, and related, the measure of cost (cQ) is not equal to marginal cost if it includes cost items that do not vary with output. Note that in the accounting approach the markup equals the profit rate when all cost items (including fixed factors like capital, and investment activities such as R&D and advertising) are included in the measure cQ.

The demand approach relies on an estimated demand curve (having data separately on prices and quantities for all products in a prespecified market) and a particular model of competition to

back out c from a first-order condition resulting from profit maximization. The production approach frees up all these restrictions on conduct and demand by computing the marginal cost of production directly from the cost minimization condition for a single variable input of production.

II.B. The Production Approach

Consider an economy with N firms, indexed by $i = 1, \ldots, N$. Firms are heterogeneous in terms of their productivity Ω_{it} and production technology $Q_{it}(.)$.[9] In each period t, firm i minimizes the contemporaneous cost of production given the production function:

$$Q_{it} = Q_{it}(\Omega_{it}, \boldsymbol{V}_{it}, K_{it}), \tag{3}$$

where $\boldsymbol{V} = (V^1, \ldots, V^J)$ is the vector of variable inputs of production (including labor, intermediate inputs, materials,...), K_{it} is the capital stock and Ω_{it} is productivity. The key assumption is that within one period (a year in our data), variable inputs frictionlessly adjust, whereas capital is subject to adjustment costs and other frictions. Because in the implementation we use information on a bundle of variable inputs and not the individual inputs, in the exposition we treat the vector $\boldsymbol{V}$ as a scalar V.[10] We consider the Lagrangian objective function associated with the firm's (conditional) cost minimization:[11]

$$\mathcal{L}(V_{it}, K_{it}, \lambda_{it}) = P^V_{it} V_{it} + r_{it} K_{it} + F_{it} - \lambda_{it}(Q(\cdot) - \overline{Q}_{it}), \tag{4}$$

where P^V is the price of the variable input, r is the user cost of capital,[12] F_{it} is the fixed cost, $Q(\cdot)$ is the technology specified in equation (3), $\overline{Q}$ is a scalar and λ is the Lagrange multiplier.

9. We derive the expression to compute markups in the most general case of firm-specific technologies, as long as the production function is twice differentiable. We subject our main empirical findings to various robustness checks precisely related to this production technology heterogeneity.

10. Of course, we can equally consider multiple inputs facing adjustment frictions—see De Loecker and Warzynski (2012) and De Loecker et al. (2016) for a discussion. We do exactly that when we use labor as the variable input in our robustness exercise.

11. The conditional statement refers to the fact that we condition on the factors of production that are chosen dynamically. For example, if capital faces adjustment costs or simply time to build, the firm chooses variable inputs to minimize cost, given the level of capital that was set in the previous period.

12. Later we will use lowercase letters to denote logs, for example, $\log(P^V) = p^V$.

We assume that variable input prices are given to the firm.[13] We consider the first-order condition with respect to the variable input V, and this is given by:

$$\frac{\partial \mathcal{L}_{it}}{\partial V_{it}} = P_{it}^{V} - \lambda_{it} \frac{\partial Q(\cdot)}{\partial V_{it}} = 0. \tag{5}$$

Multiplying all terms by $\frac{V_{it}}{Q_{it}}$ and rearranging terms yields an expression for the output elasticity of input V:

$$\theta_{it}^{v} \equiv \frac{\partial Q(\cdot)}{\partial V_{it}} \frac{V_{it}}{Q_{it}} = \frac{1}{\lambda_{it}} \frac{P_{it}^{V} V_{it}}{Q_{it}}. \tag{6}$$

The Lagrange multiplier λ is a direct measure of marginal cost (tracing out the value of the objective function as we relax the output constraint), and we define the markup as the price–marginal cost ratio $\mu = \frac{P}{\lambda}$, where P is the output price. Substituting marginal cost for the markup to price ratio, we obtain a simple expression for the markup:

$$\mu_{it} = \theta_{it}^{v} \frac{P_{it} Q_{it}}{P_{it}^{V} V_{it}}. \tag{7}$$

The expression of the markup is derived without specifying conduct or a particular demand system. Note that with this approach to markup estimation there are in principle multiple first-order conditions (of each variable input in production) that yield an expression for the markup. Regardless of which variable input of production is used, two key ingredients are needed to measure the markup: the revenue share of the variable input, $\frac{P_{it}^{V} V_{it}}{P_{it} Q_{it}}$, and the output elasticity of the variable input, θ_{it}^{v}.

The markup formula (7) derived under the production approach highlights that the marginal cost of production is

13. This approach does not preclude input providers charging a markup over marginal cost, potentially leading to double marginalization. The method for computing markups allows for arbitrary markups along the input-output table of the economy. We maintain the assumption that the input price is not a function of the input quantity demanded, through either bargaining, bulk discounting, or long-term contracts. For a formal analysis allowing for such input price feedback see De Loecker et al. (2016). In fact, the production approach can be used to recover the input-price elasticity by exploiting multiple first-order conditions across a set of variable inputs—for applications of this approach see Morlacco (2017), Mertens (2019), and Rubens (2019).

derived from a single variable input in production, without imposing any particular substitution elasticity with respect to other inputs (variable or fixed) in production or returns to scale. It is instructive to contrast it to the accounting approach introduced above: only in the case of a CRS single variable input (V) production function without fixed costs will the correct markup be measured by the sales to the variable input expenditure.

An important component of the markup formula under the production approach is therefore the output elasticity θ^v_{it}. In Appendix A, we discuss in detail the different approaches we take to measure this, and we appraise the merits and shortcomings of each approach. We distinguish between obtaining output elasticities from estimating the production function and from cost shares.

II.C. Data

To cover the longest possible period of time and to have a wide coverage of economic activity, we use data on publicly traded firms. To our knowledge, Compustat is the only data source that provides substantial coverage of firms in the private sector over a long period of time, spanning the period 1950 to 2016. While publicly traded firms are few relative to the total number of firms, because the public firms tend to be the largest firms in the economy, they account for 29% of private U.S. employment (Davis et al. 2007).

There is a serious concern that the sample of publicly traded firms is not representative of the distribution of the universe of firms. Listed firms are bigger, older, more capital-intensive, and more skill-intensive. They also involve a bigger role for multinationals. The industry mix of Compustat firms differs from that of the private sector as a whole.[14] We deal with the selection bias from studying the publicly traded firms in two ways. In Section III.D we repeat our analysis on the U.S. Censuses. For a number of sectors, we have the universe of firms. Second, we use the population weights of each sector to adjust the weights in the Compustat sample. Although we still only use publicly traded

14. There are also pronounced trends in the number and character of listed firms in recent decades. These developments are well documented in the literature. To summarize briefly, there was a huge influx of riskier, younger firms in the 1980s and 1990s (see, e.g., Fama and French 2004; Davis et al. 2007; Brown and Kapadia 2007). In something of a reversal, there has been a huge net decline in the number of U.S. listed firms since the early 2000s (see Gao, Ritter, and Zhu 2013; Doidge, Karolyi, and Stulz 2017). In the period since the mid-1990s, the average firm size has increased.

firms to calculate the markups, we account for any bias because of the sectoral composition.

The Compustat data contains information about firm-level financial statements, which allows us to rely on the so-called production approach for measuring markups. In particular, we observe measures of sales, input expenditure, capital stock information, and detailed industry activity classifications.[15] The item from the financial statement of the firm that we will use to measure the variable input is cost of goods sold (COGS). It bundles all expenses directly attributable to the production of the goods sold by the firm and includes materials and intermediate inputs, labor cost, energy, and so on.[16] In addition, we observe relevant and direct accounting information of profitability and stock market performance. The latter information is useful to verify whether our measures of markups, as discussed below, also relate to the overall evaluation of the market. Appendix Table B.1 provides basic summary statistics of the firm-level panel data used throughout the empirical analysis.

From our data, we construct a measure of the user cost of capital. We follow the standard procedure in the literature and use $r_t = (I_t - \Pi_t) + \Delta$, where I_t, Π_t, and Δ are the nominal interest rate, the inflation rate, and a depreciation rate. We use gross capital (PPEGT) that we adjust for the industry-level input price deflator (PIRIC from FRED), for the federal funds rate and for an exogenous depreciation rate and risk premium jointly that we set at 12%.[17]

Our data also have a measure of overhead, booked under selling, general and administrative expenses (SG&A). This item

15. The Compustat data have been used extensively in the literature related to issues of corporate finance, such as CEO pay, for example, Gabaix and Landier (2008), but also for questions of productivity and multinational ownership, for example, Keller and Yeaple (2009).

16. The sample does not directly report a breakdown of the expenditure on variable inputs, such as labor, intermediate inputs, electricity, and others, and therefore we prefer to rely on the reported total variable cost of production. Alternatively, we could rely on imputed intermediate inputs as in Keller and Yeaple (2009). However, that requires additional assumptions by deriving a measure of intermediate input use.

17. Below we investigate the capital share (the expenditure on capital divided by sales) and we find, not surprisingly, that this measure is quite volatile. Gross capital is a long-term measure that adjusts at a lower frequency and therefore is more subject to aggregate fluctuations. Also, in the 1970s there was a sudden drop in capital investment. Those were tumultuous financial times: inflation was high and financial frictions were considered higher.

includes selling expenses (salaries of sales personnel, advertising, rent), general operating expenses, and administration (executive salaries, general support related to the overall administration). We use SG&A to calculate total costs—not just the cost of factors of production—to measure the profits of the firms. In addition, we will consider a production technology, different from the conventional technology, where we treat overhead as a factor of production.

II.D. Censuses

As a robustness exercise and to verify the extent of selection bias in our sample of publicly traded firms, we repeat this exercise for the Economic Census. The Economic Census is administered every five years. It is composed of censuses of different sectors: a Census of Manufacturing, a Census of Retail Trade, a Census of Wholesale Trade, and so on. Within each sector, it covers the universe of employer establishments (establishments that hire workers and are not just one-person sole proprietorships); compliance is legally required.

The Census of Manufacturing contains establishment-level data on sales, in addition to very comprehensive data on inputs (the total labor wage bill, capital, materials, etc.). However, most of the other sector censuses (retail, wholesale, etc.) only contain data on establishment-level sales and wage bills, and not other nonlabor inputs. The census does not include information on overhead directly.[18] In Section III.D we analyze markups for manufacturing, retail, and wholesale. A detailed description of the census data is in Appendix B.

III. The Evolution of Markups in the U.S. economy

The bulk of our analysis is for the Compustat data where we observe firms across a wide range of sectors and time. Because we have firm-level markups, the main focus of attention is on the evolution of the distribution of markups. We first report the average markup, then detailed properties of the distribution, and finally we decompose the average markup to single out the reallocation of economic activity toward high-markup firms.

18. However, one can obtain multiple sources of information about overhead costs in the census data. We leave this for future work.

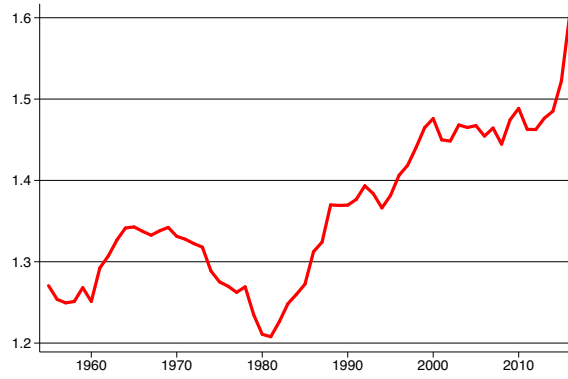


FIGURE I

Average Markups

Output elasticities θ_{st} from the estimated production function are time-varying and sector-specific (two-digit). The average is revenue weighted. The figure illustrates the evolution of the average markup from 1955 to 2016.

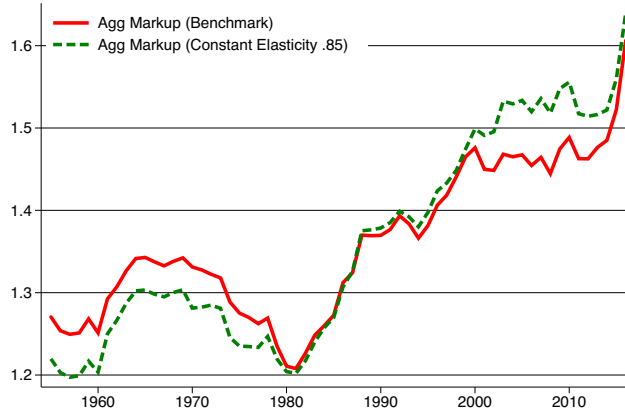
III.A. Aggregate Markups

The measure of markups in equation (7) is the product of the output elasticity θ and the inverse of the variable input's revenue share $\frac{PQ}{P^V V}$. The latter is directly measured in the firm's income statement, and we estimate the former. Our estimated output elasticities are sector- and time-specific and thus capture technological differences across sectors and time.

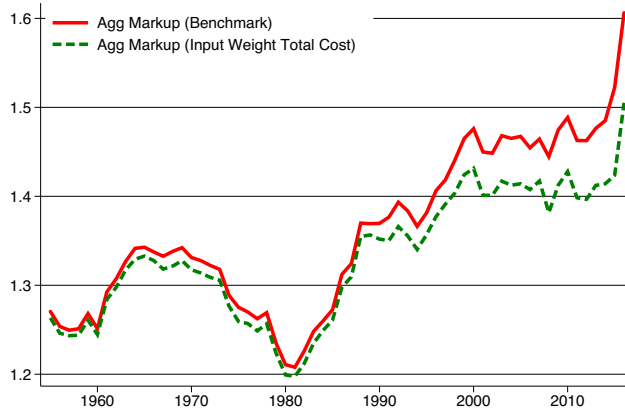
We calculate the average markup as follows:

$$\mu_t = \sum_i m_{it}\mu_{it}, \tag{8}$$

where m_{it} is the weight of each firm. In our main specification, we use the share of sales in the sample as the weight. Figure I reports the evolution of our baseline measure of average markups across the economy over time. In the beginning of the sample period, markups were relatively stable, initially slightly increasing to 1.34 in the 1960s and then decreasing to 1.21 in 1980. Since 1980 there has been a steady increase to 1.61. In 2016, the average markup charged is 61% over marginal cost, compared with 21% in 1980. In Online Appendix 5 we report a few examples of individual firms' markups.



(A) Constant elasticity



(B) Input weighted (total cost)

FIGURE II

Aggregate Markup

In broad terms, three sources can account for this rise in aggregate markups: (i) the inverse ratio of the cost share of sales, (ii) the output elasticity, (iii) the weight. To show the sensitivity of the average markups to each of these determinants, in Figure II we plot the average markup with input weights and the average markup with a fixed, time-invariant output elasticity.

When we fix the output elasticity to be time-invariant (calibrated to 0.85, the average cost share), we find that the pattern

of markups (Figure II, Panel A) is similar to that in the benchmark with estimated output elasticities. This tells us that the rise in markups is not due to the change in the estimated output elasticity, which captures technological change under our production function specification. Consistent with this evidence, we find that the output elasticities vary very little over time (see also Figure XII, Panel B, later).

Next we investigate the role of the input weight, the importance of which has first been flagged by Grassi (2017) and Edmond, Midrigan, and Xu (2019). When firms have market power, they charge higher prices and, as a result, dampen demand. With lower demand, the quantity sold and the inputs used to produce are lower. Nonetheless, revenue (price times quantity) is higher. As a result, firms with higher markups tend to have higher revenue weights relative to their input weights.

This is exactly what we see in Figure II, Panel B. The level of markups is lower throughout, and the rise is less pronounced, which indicates that the gap between inputs and sales has grown. The widening gap indicates that there is a change in the equilibrium outcome and the market structure. Moreover, as we will see in the next two sections, the widening of the markup distribution and the reallocation of sales toward high-markup firms can explain why the gap has widened.

Here we use the total cost (computed as the sum of COGS, SG&A, and rK) as the input weight and for different weighting measures the gap between the sales and the input-weighted aggregate markup is larger.[19] Because we are interested in the properties of the entire distribution of markups, we believe it is instructive to show as many different moments as possible. In particular, the gap between the input-weighted and the revenue-weighted aggregate markup informs us about the underlying mechanism, the underlying distribution, and the reallocation.

We use as our benchmark the revenue-weighted markup for the following reasons. First, a substantial portion of what is going on in the output market is reallocation (see below) of revenues

19. We revisit the weighting extensively in the robustness part of Section VI, after we have introduced the markup distribution, reallocation, and profit measures.

toward high-markup firms. We cannot capture this crucial phenomenon with input-weighted markups. The revenue-weighted markup therefore informs us about the economic mechanism and we show in a companion paper (De Loecker, Eeckhout, and Mongey 2018) that this is an important determinant in explaining the rise of market power. Second, to study market power, we link markups with profit rates (Section IV). Profit rates are traditionally aggregated with revenue weights, and consistency then calls for revenue weighting of the markup as well. Finally, revenue weighting is a common benchmark that is commonly used, most notably for widely used economic indicators, such as GDP and, in the context of market power, HHI.

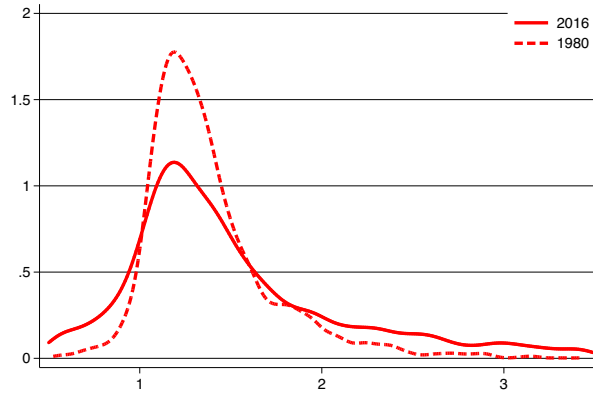
The bottom line is that our finding for the benchmark measure of aggregate markups is robust. This implies that the bulk of the action comes from the increase in the wedge of sales to COGS. The rise is not driven by technological change (changing output elasticities) and the weighting scheme informs us about the underlying mechanism where the increasing gap between the revenue-weighted and input-weighted aggregate markup tells us that firms spend less on variable inputs.

III.B. The Distribution of Markups

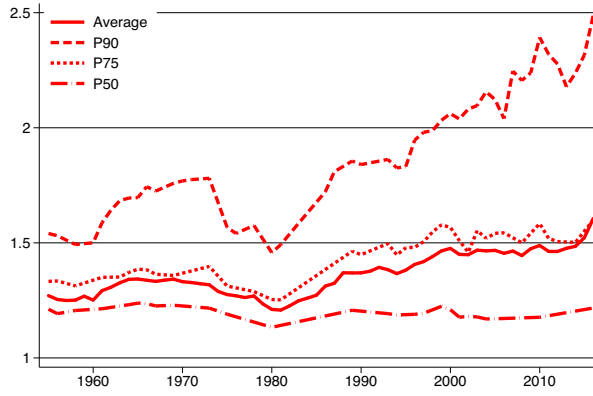
Although average markups make for a good headline, they do not fully capture the underlying distributional change in markups. The advantage of our method to calculate markups is that we obtain one for each firm, so we have a distribution of markups. A key finding is that the increase in markups is driven by a few firms, without any increase for most.

To get an idea of the evolution of the entire distribution of markups, we plot the kernel density of the unweighted markups for 1980 and 2016 (Figure III, Panel A). We find that the variance has increased and that, in particular, the upper tail has considerably fattened and become longer. It is the upper tail that drives the increase in the average markup.

Because the kernel density does not take into account the weights, we next plot the different moments of the distribution of sales-weighted markups over time (Figure III, Panel B). We rank the firms by markup; to obtain the percentiles we weight each



(A) Kernel density (unweighted)



(B) Percentiles markup distribution (revenue weight)

FIGURE III

The Distribution of Markups μ_{it}

firm by its market share in the entire sample. This makes the percentiles directly comparable to our share-weighted average. The ranking is updated each year, so the firms at the top may be different each year (later we investigate the persistence in the markup process).

The increase in the average markup comes entirely from the firms with markups in the top half of the markup distribution. The median (P50) and the percentiles below the median are

invariant over time. Most firms see no increase in markups.[20] For the higher percentiles, markups increase. For the 90th percentile in particular, the increase is sharpest. Between 1980 and 2016, it increases from 1.5 to 2.5. This indicates that the change in average markup is largely driven by a few firms that currently have much higher markups than decades ago.[21]

III.C. Reallocation of Economic Activity

The rise in the average markup is driven by a few firms at the top of the distribution. Most firms see no increase in markups, while a few firms see a large increase. We can further decompose the increase in the weighted average markup into the component that is attributable to the increase in the markup itself, and the component that is attributable to the reallocation of economic activity towards high-markup firms.

Inspection of Figure III, Panel A already shows that there is a change in the distribution of unweighted markups. The fatter tail is evidence that more firms have higher markups. Even if the distribution of unweighted markups had remained unchanged, the weighted aggregate markup could have gone up if the firms with higher markups now obtain a higher share of the market. This reallocation of economic activity toward higher-markup firms is important to understand the implication that market power has on the concentration of economic activity in the hands of a few dominant firms. Though not in all, in most theories of market power, firms that have higher market power also increase their market share (in the Cournot model in particular, the market share is a sufficient statistic of market power).

Because the change in aggregate markups is a combination of the rise in unweighted markups and a reallocation of economic activity, we decompose the average markup at the firm level as

20. Because the distribution is revenue-weighted and the larger firms tend to have higher markups, this implies that the vast majority of firms see no rise in markups.

21. This is consistent with the evidence in Kehrig (2011). He studies the cyclicality of productivity and finds that the dispersion in TFPR is increasing, especially in the upper tail, where TFPR captures both markups and cost-side heterogeneity. In Appendix E we further explore the distributional change by modeling the markup (as well as sales and employment) as an autoregressive process and confirm the rise in the standard deviation.

follows:

$$\Delta\mu_t = \underbrace{\sum_i m_{i,t-1}\Delta\mu_{it}}_{\Delta\text{within}} + \underbrace{\underbrace{\sum_i \tilde{\mu}_{i,t-1}\Delta m_{i,t}}_{\Delta\text{market share}} + \underbrace{\sum_i \Delta\mu_{i,t}\Delta m_{i,t}}_{\Delta\text{cross term}}}_{\Delta\text{reallocation}}$$

$$+ \underbrace{\sum_{i\in\text{Entry}} \tilde{\mu}_{i,t}m_{i,t} - \sum_{i\in\text{Exit}} \tilde{\mu}_{i,t-1}m_{i,t-1}}_{\text{net entry}}, \tag{9}$$

where $\tilde{\mu}_{it} = \mu_{it} - \mu_{t-1}$ and $\tilde{\mu}_{it-1} = \mu_{it-1} - \mu_{t-1}$.[22]

We apply the insights from the productivity-decomposition literature, and while this decomposition appears very similar to that in equation (10), it is different, first because it has one additional term, and second because its interpretation is very different. There is an additional term here because there is entry and exit of firms, whereas in the sectoral decomposition the number of sectors is fixed.[23] The interpretation also differs. Following Haltiwanger (1997), we consider a theoretical counterfactual where the Δwithin term measures the average change that is merely due to a change in the markup, while keeping the market shares unchanged from last period. Instead, the Δmarket share term measures the change due to an increase in market share while keeping the markup fixed. If this term is increasing, it captures the fact that firms with higher markups now have a higher market share, and hence there is an increase in the weight of the high-markup firms. This in turn raises the average markup without raising the markup itself. The Δcross term measures the joint change in markups and market share. We denote by Δreallocation the joint effect of Δmarket share + Δcross term.[24] Finally, the new last term measures the effect of entry and exit on markups. This captures the change in the composition of firms in the market. If

22. We demean the (lagged) markups by the appropriate aggregate (share-weighted) level, to correctly identify the role of the reallocation term—see Haltiwanger (1997) for more discussion.

23. Entry and exit in the data of publicly traded firms comprise entry and exit in the database. This consists of firms listing and delisting, as well as merger and acquisition activity.

24. The Δcross term is virtually 0 in the experiments we perform.

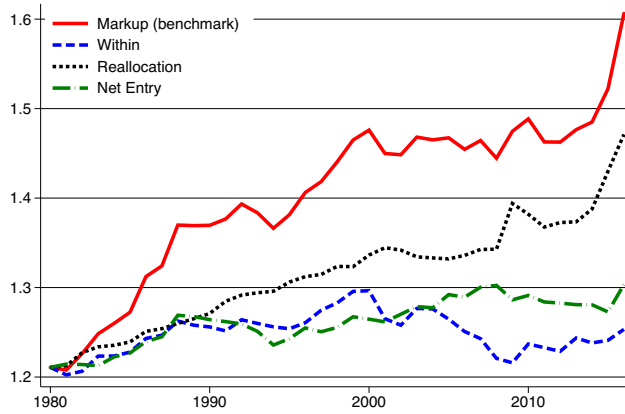


FIGURE IV

Decomposition of Markup Growth at the Firm Level

the entering firms have higher markups than the exiting firms, for example, this term will be positive.

We perform this decomposition across firms in the entire economy. To best present this decomposition, Figure IV plots the average markup (solid line in print; solid red in color version online; color version online), as well as three counterfactual experiments based on the decomposition starting in 1980. We set the initial level to 1980 and then cumulatively add the changes of each component term in equation (9).[25]

The first experiment (long dashed line in print; solid blue in color version online) shows the evolution of the average markup as if there was only component Δwithin and all other components were 0. This shows that the rise in average markups in the 1980s and 1990s from 1.21 to 1.3 in 2000 is about one-third of the total increase from 1.21 to 1.47. From 2000 onward, this term decreases and picks up again after the Great Recession. The change in the average markup is also evident from Figure III, Panel A, where we see an increase in the upper right tail.

The second experiment (short dashed line in print; solid black in color version online) shows the path of the markup if the only change had been due to Δreallocation. All markups

25. In Online Appendix Section 4, we tabulate the measured yearly changes of each of the four components for all years between 1955 and 2016. The cumulative representation in Figure IV shows decomposition of the change in markups in a more concise way.

remain unchanged from the previous period, and we apply only the change in the market shares. The plot shows that accumulated over the whole time period, reallocation accounts for about two-thirds of the change in the weighted markup. The main takeaway here is that there are two forces at work. On the one hand, the markup (the within term) increases, which is an indication of the change in pricing power of firms. In De Loecker, Eeckhout, and Mongey (2018), we show that this can be due to a change in the market structure (less competition) or due to technological change (bigger spread in firm productivity). On the other hand, there is also a reallocation of sales activity away from low-markup firms toward high-markup firms (the reallocation term). This is entirely consistent with a model of imperfect competition where firms with higher markups also attract a higher market share. This reallocation effect is in accordance with the findings in Autor et al. (2020) and Hartman-Glaser, Lustig, and Zhang (2016), who establish that large firms have grown in size relative to small firms, and those firms tend to operate in more concentrated markets. While we find that the reallocation term is important, it is not the only force at work. Unweighted markups have gone up (measured by the Δwithin term and visualized by the density of markups in Figure III, Panel A, especially in the upper tail), which is an important force behind the rise in market power. In a general equilibrium model with input-output linkages, Baqaee and Farhi (2020) find a similar decomposition of the within and the reallocation component.

The third experiment (dash and dot line in print; solid green in color version online) shows the evolution of markups if the only change was net entry of firms. The net entry component rises early on and is more or less constant afterward, indicating that the rise in markup is not exclusively driven by the changing composition of firms in the sample. The net entry component can simply be driven by the fact that the panel of firms is not balanced and more firms enter than exit. In part, it can also be driven by mergers and acquisitions. Consider two firms that merge. If their joint market share is unchanged but they now charge higher markups, then the net entry term will be positive. Or it could be driven by the fact that the net entry accounts for a higher market share than the sum of the individual premerger shares.

In summary, the rise in aggregate markups is driven in part by a change in the markup distribution itself, by a reallocation from low-markup firms to high-markup firms, and by some net

TABLE I
SECTORAL DECOMPOSITION OF 10-YEAR CHANGE IN MARKUP

	Markup	Δmarkup	Δwithin	Δbetween	Δcross
1966	1.337	0.083	0.057	−0.017	0.041
1976	1.270	−0.067	−0.055	0.002	−0.014
1986	1.312	0.042	0.035	0.010	−0.003
1996	1.406	0.094	0.098	0.004	−0.008
2006	1.455	0.049	0.046	0.007	−0.005
2016	1.610	0.154	0.133	0.014	0.007

entry. In the first decade of the sample, the 1980s, all three forces are equally at work. But by the end of the period, reallocation dominates. Cumulatively over the whole period, reallocation accounts for two-thirds of the rise in markups.

The decomposition exercise implies that the reallocation component captures movements of firms across all sectors. In Online Appendix 4 we perform the same decomposition for each of the broad sectors of the economy, where reallocation of economic activity is measured within sector.

In contrast to the firm-level decomposition (economy-wide and within sector), we also analyze the decomposition of the rise of markups by firm size at the sectoral level, that is, within and between sectors. Is the increase in markup over time due to a change of markup at the industry level (Δwithin), due to a change in the composition of the firms—there are more firms with a high markup—(Δbetween), or due to the joint change in markup and the firm composition (Δcross term)? This can be expressed in the following formula:

$$\Delta\mu_t = \underbrace{\sum_s m_{s,t-1}\Delta\mu_{st}}_{\Delta\text{within}} + \underbrace{\sum_s \mu_{s,t-1}\Delta m_{s,t}}_{\Delta\text{between}} + \underbrace{\sum_s \Delta\mu_{s,t}\Delta m_{s,t}}_{\Delta\text{cross term}}. \tag{10}$$

We consider the change over 10-year periods starting in 1956 in Table I.[26] The decomposition shows that the change in markup is mainly driven by the change within industry. Most of the Δmarkup is driven by Δwithin. There is some change in the composition between industries, but that is relatively minor compared

26. The decomposition for the three- and four-digit industry classification is reported in the Online Appendix.

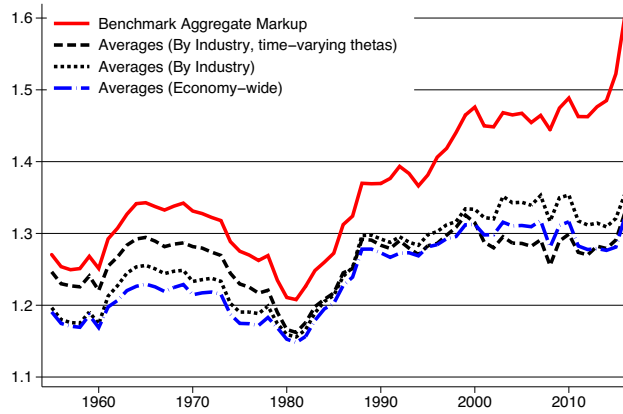


FIGURE V

Using Industry and Economy-Wide Averages versus Aggregating Microdata

to the within industry change. The change due to reallocation, the joint effect, is mostly small.

In sharp contrast with the firm-level decomposition where most of the increase is due to reallocation between firms, the sectoral decomposition shows that most of the increase in markups occurs within all sectors, not between sectors. This is an important and unexpected discovery. Intuitively, we would expect that certain sectors, such as technology, would see a much bigger increase in the markup. But as the sector-specific markups in Online Appendix Figure 12.1 illustrate, there are no sectors that systematically have higher market power. This confirms that the increase in market power occurs in all sectors and industries.

Further evidence that most of the rise in market power occurs within industry comes from comparison of our results with those based on aggregate data (industry-level or economy-wide). Using national accounts data by sector, Hall (2018), extending his original work (Hall 1988), finds a rise in market power but only by about 20 points, half of the increase we find with firm-level data.

To investigate where the discrepancy when using aggregate data comes from, we use our firm-level data and aggregate them at the industry level. In Figure V, we plot our benchmark aggregate markup together with three series of industry averages, based on our firm-level data, summed up to industry averages: one where we treat the entire economy as one industry (dashed-dotted line in

print; solid blue in color version online), one where we aggregate at the industry level with constant elasticities (short-dashed line in print; long-dashed black line in color version online), and one where we aggregate at the industry level and use the estimated, time-varying and industry-specific elasticities (long-dashed line in print; solid black line in color version online).

The three series with averages look similar. The average markups are below our benchmark, and it grows at half the rate. The increase between 1980 and 2016 is from 1.15 to 1.35 approximately, by about 20 points, as in Hall (2018). This clearly establishes that a substantial part of the increase occurs within industry and that some of that change is lost when taking averages. To see how that can occur, consider the following comparison. To make the comparison as transparent as possible, we abstract from any technological change or sectoral heterogeneity in output elasticities and simply keep θ constant throughout. We compare our aggregate markup with the one obtained using aggregate data:

$$\sum_i m_{it} \frac{S_{it}}{P_{it}^V V_{it}} \neq \frac{\sum_i S_{it}}{\sum_i P_{it}^V V_{it}}. \tag{11}$$

The reason the two objects are not equal to each other is because of the heterogeneity in markups across firms.[27] Aggregation of a nonlinear function (Jensen's inequality) leads to different outcomes. This is the case for any cross-section, but importantly with the reported increasing skewness in the underlying markup distribution, this difference becomes larger over time. The widening gap between the micro and the macro ratios is simple economics: if market share is reallocating toward the higher-markup firms, this reinforces the process of increased skewness, due to the increased correlation of markups and market share (in a given industry or in the entire economy depending on the focus).

It is clear from the figure that the aggregate-based series trend up but to a much lesser extent, and this is to be expected given the increased dispersion. This tells us that the dispersion and skewness of the distribution have increased over time. Much of the rise that we observe in the average markup disappears once we use industry or economy-wide averages. This tells us that

27. With identical firms, the market share is $m_{ij} = N^{-1}$ and both ratios are identical.

most of the heterogeneity in markups is within industry and that the reallocation of market shares (see also below) occurs mainly within industries.

III.D. Results from the U.S. Censuses

The data on publicly traded firms suffers from selection. So far, what we have analyzed cannot be generalized to the entire U.S. economy. The publicly traded firms tend to be large, and the number of firms (fewer than 10,000) is small relative to the approximately 6 million firms in the economy. Moreover, entry and exit in the sample of publicly traded firms is nonrandom. Even though the shares of GDP and of employment are large (because the firms are large), we want to find out whether our results are representative for the entire economy.

To that end, we repeat the exercise for the censuses in different industries. The advantage of the censuses is that they represent the universe of firms within a sector and are therefore representative of the whole economy in that sector. We focus on three censuses: manufacturing (NAICS codes 31-32-33), wholesale (NAICS code 42), and retail (NAICS codes 44-45). We provide more detail on the sample construction and measurement of the key variables in Appendix B.

The measurement of markups in the census data relies on the framework outlined in Section II. The implementation, however, differs because we do not observe the same detailed information as in Compustat regarding a firm's balance sheet and income and loss statement, with the exception of the Census of Manufacturing, for which we do observe most of the traditional production and cost variables. The analysis of the manufacturing sector will therefore closely track the analysis applied to the universe of Compustat firms. There remains one big difference: to our knowledge there is no analogue to the reporting of SG&A (or overhead cost) in the census data.[28]

With the exception of the Census of Manufacturing data, we only observe the wage bill and sales consistently across plants and time. This implies that output elasticities cannot be measured or estimated due to the limited information on costs. For manufacturing, where there is more detailed reporting of costs, we use the

28. Some components such as marketing and advertising costs are in principle recorded, but items such as brand value, research and development, and executive compensation packages, are not.

industry-time specific cost shares as measures for output elasticities. For retail and wholesale, we cannot impute the cost shares. Instead, we use the sector- and time-specific output elasticities that we estimated from the publicly traded firms.

In the Census of Manufacturing, we use the cost shares to construct the output elasticity of any variable input (labor and materials) at the four-digit NAICS industry level (denoted by n) by census year.[29] This leads to the standard recovery of the output elasticity for the variable input:

$$\theta_{nt}^{V} = N_{nt}^{-1} \sum_{j \in n} \frac{P_{jt}^{V} V_{jt}}{P_{jt}^{V} V_{jt} + r_{nt} K_{jt}}, \tag{12}$$

where j denotes a plant active in industry n, in this case a unique four-digit NAICS code.[30] For manufacturing we can use information on materials and on the wage bill for the variable input V. This allows us to check the robustness of our findings. For the other censuses, we only observe the wage bill. In the absence of information on cost shares, we infer the output elasticities of labor using the cost-share approach in Compustat. In particular for each two-digit NAICS sector (s), we compute the median labor cost share, by year, for the sample of active firms, as in Section VI.A.

Finally, we aggregate the plant-level markups to obtain firm-level markups, the ultimate object of interest in this analysis. This also makes our results consistent with the analysis performed for the Compustat sample.[31] More specifically, we compute markups at the plant level and aggregate to the firm level using plant-level revenue shares. The sector-specific aggregate markup is computed as before, using a firm's share in total sectoral sales.

29. These are made available by Foster, Haltiwanger, and Syverson (2008) and are accessed through the census file. Alternatively the output elasticities can be obtained by estimating the 86 distinct production functions using an approach as outlined in Section II. We opted to rely on the cost-share approach to minimize the impact of measurement error and imputed data in obtaining reliable estimates of the output elasticity. See Syverson (2004) and Nishida et al. (2017) for a discussion of these issues.

30. We remind the reader that there are four distinct levels of aggregation in our analysis: plants (j), firms (i), industries (n, i.e., four-digit NAICS), and sectors (s, i.e., two-digit NAICS).

31. In the case of the wholesale and retail sectors, the output elasticity is measured at the level of the two-digit NAICS code: NAICS 42 and 44-45 combined, respectively.

Figure VI reports the weighted average (left panels) for each of the three censuses, as well as the percentiles of the markup distribution (right panels), weighted by sales (the equivalent of Figure III, Panel B). With data only in five-year intervals, the patterns are obviously less detailed.

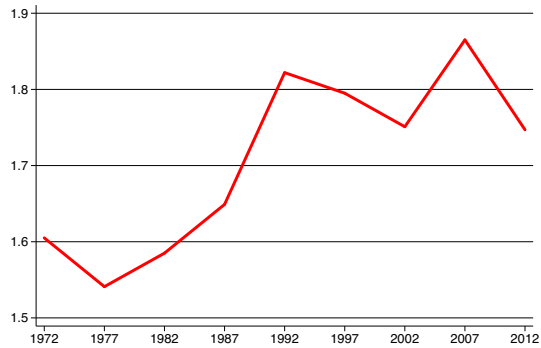
Starting with Manufacturing (Figure VI, Panels A and B), we see average markups that start to increase from 1977 onward, from around 1.55 up to around 1.8. This pattern mirrors what we find in the whole sample of publicly traded firms and in the publicly traded firms in manufacturing.

We also calculate the markup using materials as the variable input, instead of employment, and we find a very similar pattern. In the Compustat sample, we cannot separate the labor and material expenditures, instead we have to rely on the bundle COGS. The results indicate that all three series (Compustat COGS-based, census labor-based, and census materials-based) indicate the same pattern of rising aggregate markups.[32]

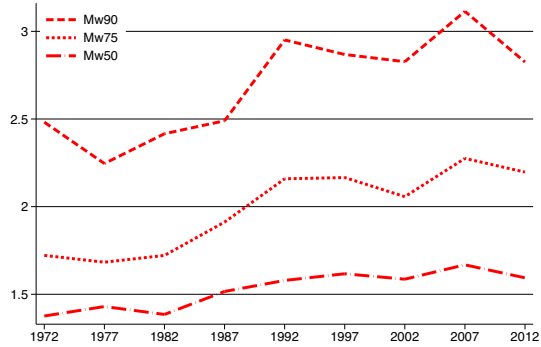
Like for the publicly traded firms, the pattern in retail (Figure VI, Panels C and D) until 2002 is flat or only slightly increasing. This is the case also for the percentiles. There is instead a sharp increase of the weighted average in 2012 that we do not observe in the publicly traded firms.

The figures for wholesale are again in line with the series obtained from our analysis in the Compustat sample. We observe a continuous decline in the aggregate markup until 2002, after which we see an increase of about 15 percentage points in the markup over the course of 10 years. The percentiles highlight again that the rise is concentrated at the top of the (weighted) markup distribution. In contrast to the results for the manufacturing and retail census, we could not rely on reliable labor cost shares to approximate the time-specific output elasticity. We describe the procedure and compare the results to reported (aggregate) profit margins in Online Appendix Section 18, but the same message holds: the time-series markup pattern is dominated by the dynamics in the sales-to-expenditure (here the wage bill) ratio, and the output elasticity mostly affects the level.

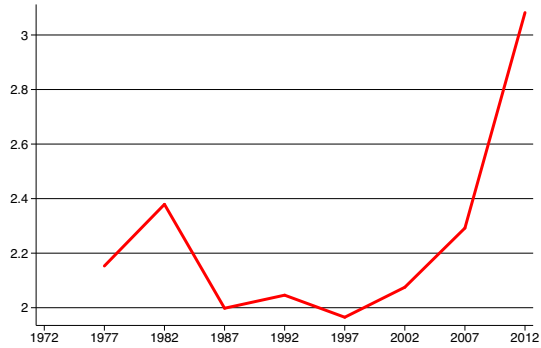
32. For more details on the use of multiple variable inputs in the manufacturing sector, see Online Appendix Section 17.



(A) Manufacturing: average



(B) Manufacturing: percentiles

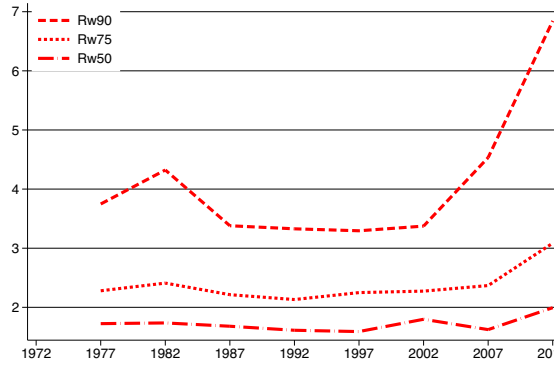


(C) Retail: average

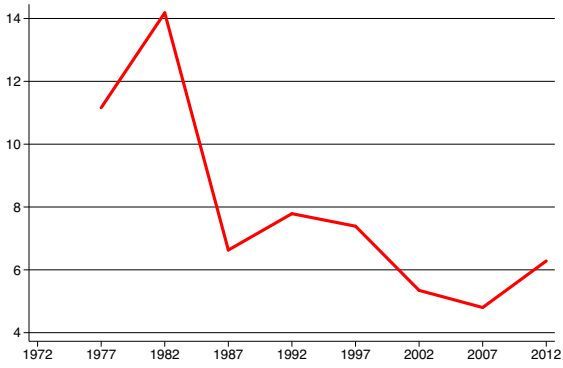
FIGURE VI

Markups in the U.S. Censuses: Manufacturing, Retail, and Wholesale

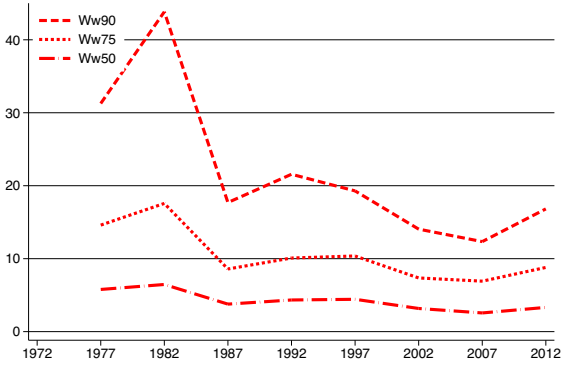
The variable input is employment. Averages and percentiles are revenue weighted. Manufacturing firm-level markups rely on the industry-specific cost shares. Retail trade relies on the output elasticities computed in the Compustat sample. Wholesale relies on a calibrated output elasticity.



(D) Retail: percentiles



(E) Wholesale: average



(F) Wholesale: percentiles

FIGURE VI
(CONTINUED)

IV. Market Power and Profitability

The documented rise in markups does not necessarily imply that firms have more market power and therefore higher economic profits. In fact, increasing markups can come from a variety of reasons that are not associated with a decline in aggregate welfare.[33] For example, a decrease in marginal costs, an increase in fixed costs or innovation, an increase in demand or in its elasticity, a change in the market structure, or new product varieties all lead to increasing markups without necessarily implying higher profits.

Although the textbook definition of market power is the case whereby a firm can command a price above the marginal cost of production (markup), any conclusions regarding whether market power increased will greatly depend on the pattern of overhead costs, or any other factor affecting the cost structure of firms (like innovation activities such as R&D). Therefore, before we can conclude whether the higher markups are associated with market power, we need to analyze profits. In the absence of detailed data, the mapping from markups to market power (and therefore welfare) can only be done through a particular model of the economy.

With the accounting data available, we assume that we can observe profits as the wedge between sales and all variable and fixed costs (including innovation, advertising, and others). In what follows, we consider higher market power a situation whereby a firm can generate higher profits.[34]

Key here is the evolution of overhead and capital as a share of expenditure. If those have increased and markups have increased at the same rate, then the higher markups are charged only to recover the higher overhead costs and capital investment. In Figure VII we plot the evolution of overhead and capital as a share of total costs.[35] In our data, the capital share has been fairly constant, in line with the findings by Barkai (2017). Instead,

33. It does, however, potentially generate distributional implications.

34. Profits do not necessarily derive exclusively from market power. There could be capital market imperfections that constrain investment and lead to higher profits. However, in a model with both market power and financial frictions, Cooper and Ejarque (2003) find that profitability is explained entirely by market power and not by financial frictions. In this article, however, we abstract away from such frictions.

35. The series are obtained by taking the ratio between total overhead (capital) and aggregate total cost; alternatively this can be interpreted as weighting each firm with its share of total cost in aggregate total cost.

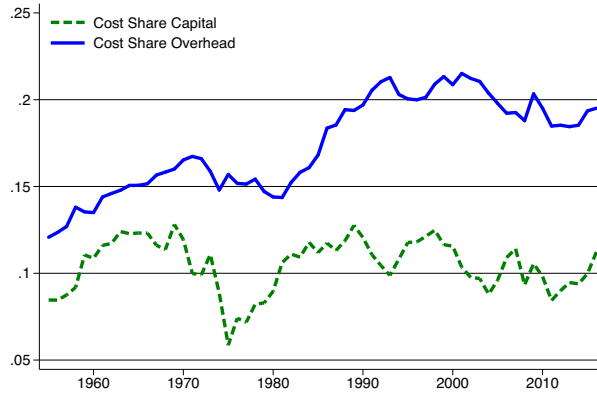


FIGURE VII

Aggregate Overhead and Capital Cost Shares of Total Cost

overhead as a share of total expenditure has seen an increase. The rise in overhead costs thus requires us to analyze profits to conclude whether the rise in markups is associated with a rise in market power.

We proceed in two steps. First, we relate markups to recorded profits at the firm level and contrast the observed markups to counterfactual markups generated by a zero-profit condition. Second, we consider aggregate profits and ask whether these are consistent with our estimates of firm-specific markups and recorded fixed costs.

IV.A. Markups and Profits at the Firm Level

To calculate profits, we use the markup measure and properly account for all costs, including the overhead (or fixed) costs and the expenditure on capital. We then interpret this profit rate as a measure of market power.

Let $\Pi_i = S_{it} - P_t^V V_{it} - r_t K_{it} - P_t^X X_{it}$ denote net profits, where $P_t^X X_{it} = F_{it}$ denotes expenditure on overhead as measured by SG&A and is equal to the fixed cost.[36] Then the net profit rate

36. We distinguish the notation depending on whether overhead is interpreted as a fixed cost (F) or as an input of production with quantity X and unit price p^X.

$\pi_{it} = \frac{\Pi_{it}}{S_{it}}$ can be written as:

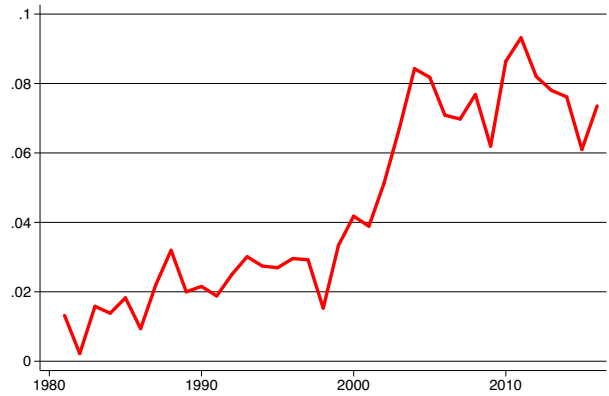
$$\pi_{it} = 1 - \frac{\theta_{st}}{\mu_{it}} - \frac{r_t K_{it}}{S_{it}} - \frac{P_t^X X_{it}}{S_{it}}, \tag{13}$$

where we have substituted the expenditure on variable inputs as a share of sales with the output elasticity over the markup, from equation (7). This measure of the profit share is different from the accounting profits because it uses a measure of capital obtained from the balance sheet, not the income statement. With adjustment frictions, the accounting measure does not adequately reflect the expenditure on capital. Note also that our measure of profits incorporates the output elasticity of the production technology, which takes into account that the variable factors of production V adjust while the fixed factors do not.

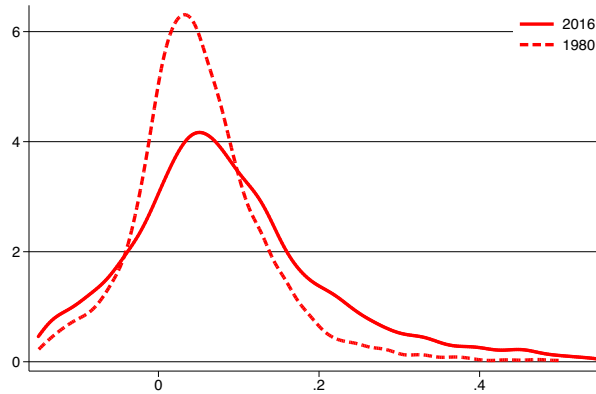
Figure VIII, Panel A plots the average revenue-weighted profit rate for the data in our sample. We find that profits have gone up by about 7 percentage points between 1980 and 2016.[37] Underlying the rise in profits is the increase in the upper tail of the profit distribution. In Figure VIII, Panel B we plot the kernel density of the unweighted profit rate distribution in 1980 and 2016. The rise in average profit rate is nearly exclusively driven by the increase in the upper percentiles of the profit distribution. More firms have extremely high profit rates of 15% and higher. Consistent with the results on markups, the average profit rate increase is driven in part by the reallocation of economic activity toward high-profit, dominant firms.

Our measure of the profit rate is the firm profits as a share of sales, which effectively scales those profits by the firm size as measured by its revenue. From an investment viewpoint, we may want to measure the return on assets. The return on assets is calculated as the firm profits divided by its assets. We define profits by sales minus all costs, COGS, SG&A, and the expenditure on capital. Because the expenditure on capital is included, our measure of return on assets is the return over and above r, which

37. Our measure of profits was also high in the mid-1970s (see Online Appendix Figure 8.1), but that is entirely driven by the drop in capital expenditure during a period of high inflation. Once we consider gross profits, without subtracting the expenditure on capital, there is no such spike in the profit rate in the 1970s.



(A) Average profit rate (revenue weighted)



(B) Kernel density profit rate (unweighted)

FIGURE VIII

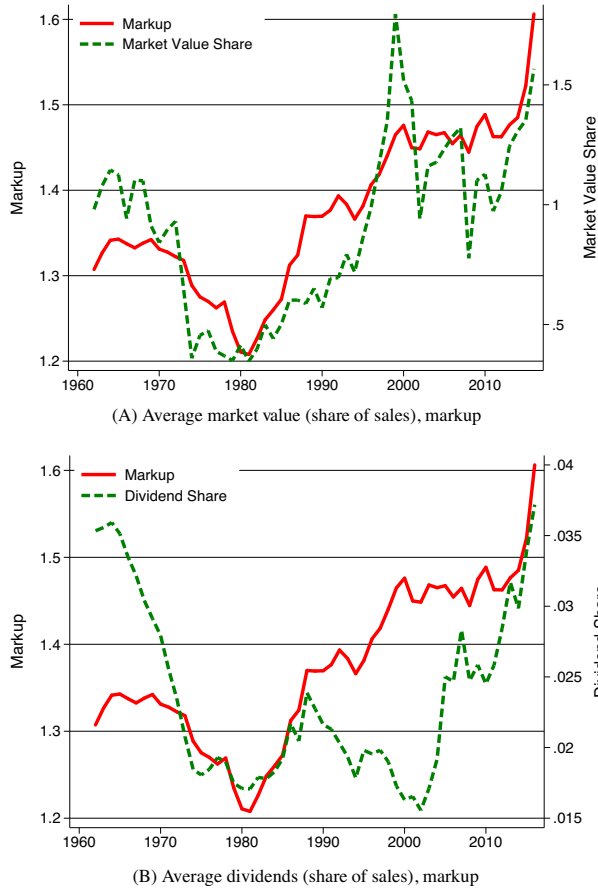
Average Profit Rate and Profit Rate Distribution

includes the inflation-adjusted risk-free rate, as well as an adjustment for depreciation and risk. Therefore, it is the excess return on assets. We plot this in Online Appendix Figure 9.1a together with our baseline profit rate. The return on assets is remarkably similar to the profit rate, with an increase starting in 1980 and rising from around 1% to around 8% in 2016. This average return on assets is weighted by the capital of each firm. When we weight it by the sales of each firm (Online Appendix Figure 9.1b), then the average return on assets is higher and also rising faster. Firms

with high sales have higher returns on assets, and the large firms have seen a bigger rise in their returns.

All this seems to suggest that at least based on the flows reported in the accounting data, starting in 1980 there has been an increase in the profitability of firms, and therefore an increase in market power. Note that the profit rate we have reported accounts for the increase in contemporaneous overhead costs as measured by SG&A. Of course, some costs may have been incurred earlier. Still, it is not clear what those startup costs may be as they are not booked in the firms' accounts, and firms have incentives to book as many costs as possible to reduce corporate taxes on profits. The only possibility is that those startup costs were incurred before the firms were observed in our data. As a result, profits based on contemporaneous costs may therefore be overstated. What the data are indicating, however, is that if such costs are incurred earlier, there must be an increase in those startup costs as a share of the sales of a firm since 1980. With free entry and hence zero ex ante expected profits, what we expect is that over the past four decades, the unmeasured startup cost as a share of future sales has gone up from 1% of sales to 8% of sales (roughly from 2% of value added to 16%). Some of those costs could be R&D costs that were incurred before the firms were observed in our data. We turn to the impact of recorded R&D costs below.

The flow of profits may not be the best measure of profitability of the firm, because it mixes up the firm's result with investment decisions. To that effect, we consider a measure of profitability based on what firms generate as a return to their shareholders. For that we have two measures: (i) the market value (or market capitalization), and (ii) dividends. Our second measure, dividends, is the return an investor receives on holding equity in the firm. Of course, dividends may vary for reasons that have nothing to do with the actual flow of profits. In particular, they will be closely related to the investment opportunities the firm has. Still, over a long enough horizon and averaging out over a large number of firms, we would expect that dividends are a good indicator of profits. Our first measure, market value, is essentially the discounted sum of dividends, since a shareholder who sells shares in a firm gives up the opportunity value of receiving the indefinite stream of dividend payments. In contrast to the actual dividends, the market price is more a measure of future expected profits, not just contemporaneous



(A) Average market value (share of sales), markup

(B) Average dividends (share of sales), markup

FIGURE IX

Market Value and Dividends

profits, since it takes into account the flow of all expected future dividends.

Figure IX, Panel A shows the evolution of the market value as a share of sales, averaged by the sales share in the entire economy: $\sum_i \frac{S_i}{\sum_i S_i} \frac{MktVal_i}{S_i} = \frac{\sum_i MktVal_i}{\sum_i S_i}$. Unlike standard composite indices of stock market values like the S&P 500, this measure is a "rate" that can be interpreted in conjunction with the profit rate π (profits as a share of sales) from our model. As such, first, it is not

affected by inflation[38] and second, this measure is independent of the size of firms or the composition of firms because it is normalized by sales. For example, even if there are 500 firms in the index, the index will artificially grow when firms become larger, for example, due to mergers.[39]

If the flow of profits and dividends as a share of sales were constant, then the market value that reflects the discounted stream of dividends would be constant as a share of sales. This is clearly not the case. Market value as a share of sales rises from less than 50% in 1980 to over 150% in 2016 (Figure IX, Panel A, right scale). A similar pattern arises for dividends, where dividends as a share of sales increases from 1.7% in 1980 to over 3.5% in 2016 (Figure IX, Panel B).[40]

This is not just an artifact of the aggregate data. At the individual firm level, firms with higher markups also have higher market values and dividends. In Table II we report the regression results.[41] Not surprisingly, contemporaneous firm-level markups are correlated with both market value and dividends. For all specifications, the coefficient is highly significant (even in the presence of firm fixed effects, see columns (4) and (8)). At the firm level, this is consistent with the fact that higher markups reflect higher profits and therefore higher dividends and market values.

Based on the evidence from the firm's fixed overhead as measured by SG&A and the resulting profits and by market value and dividends, we find evidence that the rise in markups is associated with the rise in market power.

38. The increase of the Dow Jones in the 1970s, for example, is misleading because during that period of high inflation, once adjusted for inflation, the real index is actually decreasing.

39. Interpreting the market valuation of a firm as the discounted stream of profits obviously imposes a set of assumptions. The most important one is the fact that the discount rate has remained constant over the period. We know that the risk-free rate has decreased, especially since the 1990s. While the interest rate is not the discount rate, a preference parameter, it is quite feasible that the risk-free rate affects the valuation of stocks. And of course, changes in legislation affect tax incentives and therefore firm valuation (see Smith et al. 2017).

40. Note that the reason for the decline in dividends in the 1990s and sudden increase since the early 2000s is due to tax incentives for firms to issue dividends. Until the 2003 tax reform, dividends were taxed at the individual's income tax rate, and at 15% thereafter.

41. In Online Appendix Table 19.1 we report the same regressions for our markups estimated with the technology that includes overhead as a factor of production, PF2. The coefficients are very similar.

TABLE II

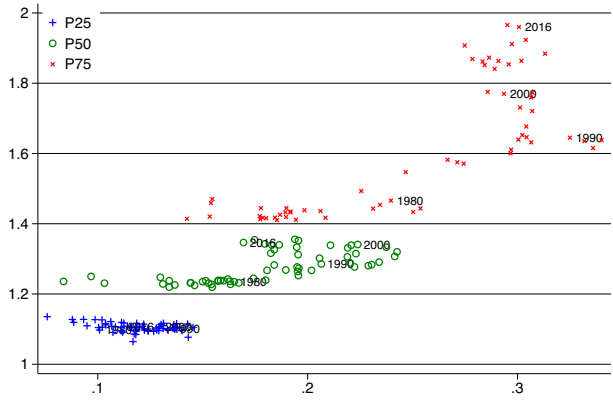
FIRM-LEVEL REGRESSIONS: MARKET VALUES AND DIVIDENDS ON MARKUPS

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	$\ln\left(\frac{\text{market value}}{\text{sales}}\right)$				ln(market value)			
ln(markup)	0.71	0.64	0.56	0.17	0.71	0.65	0.58	0.27
	(0.03)	(0.02)	(0.02)	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)
ln(sales)					0.81	0.81	0.83	0.68
					(0.00)	(0.00)	(0.00)	(0.01)
Year fixed effects		Y	Y	Y		Y	Y	Y
Sector fixed effects			Y				Y	
Firm fixed effects				Y				Y
R^2	0.05	0.13	0.21	0.68	0.68	0.71	0.73	0.89
	$\ln\left(\frac{\text{dividends}}{\text{sales}}\right)$				ln(dividends)			
ln(markup)	1.05	0.97	0.80	0.26	1.03	0.93	0.78	0.26
	(0.04)	(0.03)	(0.04)	(0.05)	(0.04)	(0.04)	(0.04)	(0.05)
ln(sales)					0.94	0.92	0.93	0.76
					(0.01)	(0.01)	(0.01)	(0.02)
Year fixed effects		Y	Y	Y		Y	Y	Y
Sector fixed effects			Y				Y	
Firm fixed effects				Y				Y
R^2	0.06	0.11	0.17	0.70	0.66	0.68	0.70	0.89

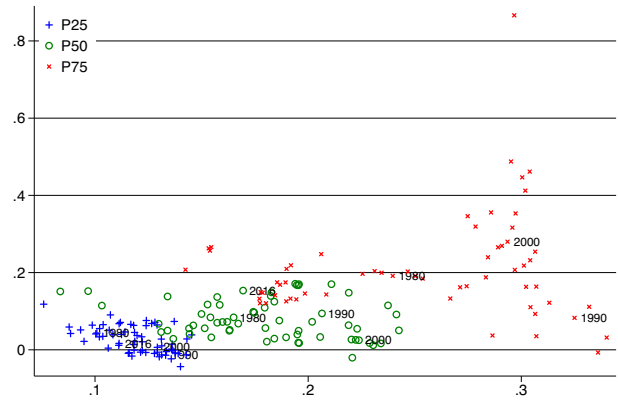
Note: Standard errors clustered by firm are in brackets.

To complete this section, we investigate the relation between profits, markups, and overhead costs (SG&A). In Figure X, Panel A we plot the relation between the share of sales of SG&A and the markup for different percentiles in the (unweighted) markup distribution. This shows that the firms with a higher SG&A share of sales have higher markups. For a given year, the higher percentiles in the distribution of markups have higher overhead shares. This is as expected in a competitive economy: higher prices relative to marginal cost are required to offset the overhead and avoid making losses. In addition, over time, the overhead share is increasing which automatically implies that the markup increases, even in a competitive economy. Note that if we plot the markup against the share of COGS in sales, then by construction this relation is downward sloping, indicating that unlike SG&A, COGS is a variable input.

Now we want to evaluate whether the increase in markups that we observe is merely to offset the rise in overhead. To that effect, we calculate a fictitious markup, denoted by $\mu^\star$, that corresponds to zero profits. We obtain that markup from setting profits



(A) Markups μ_{it} by SG&A share



(B) Excess markup $\mu_{it} - \mu_{it}^{\star}$ by SG&A share

FIGURE X

Markup, Excess Markup, and SG&A Share (Markup PF2)

π_{it} to zero in equation (15) and solving for μ:

$$\mu_{it}^{\star} = \frac{\theta_{st}}{1 - \frac{r_t K_{it}}{S_{it}} - \frac{P_t^X X_{it}}{S_{it}}}. \tag{14}$$

This zero-profit markup is a weak upper bound, however, and the true zero-profit markup is weakly lower (provided there are no costs in addition to COGS, SG&A, and capital). This is because we do not know what sales S_{it} would be under competition. To predict

sales under perfect competition, we need to know the properties of demand. Only in the case of unit elasticity demand will sales be invariant for different markups. In all other cases, however, sales under perfect competition will be lower than when there is market power. This is due to the fact that firms are charging higher prices only if the marginal revenue is positive, which by definition necessarily implies higher sales for higher markups. Therefore, sales under perfect competition ($S_{it}^{\star}$) will be weakly lower than under market power. Under our assumption that in the short run K_{it} and X_{it} are not variable, the expression in equation (14) where we use S_{it} instead of $S_{it}^{\star}$ is weakly higher than the true zero-profit markup.

In Figure X, Panel B we also plot $\mu_{it} - \mu_{it}^{\star}$ for different percentiles in the markup distribution. Because $\mu_{it}^{\star}$ is the upper bound of the zero-profit markup, the gap between the actual markup and $\mu_{it}^{\star}$ indicates the extent of the excess markup, over and above the markup that arises under perfect competition. We see that the excess markup is highest for the highest percentiles of the markup distribution, where incidentally the SG&A share is the highest as well. High-overhead firms have high markups but also high excess markups; and this became stronger over time (the excess markup rose from about 0.2 in 1980 to about 0.6 in 2016).

When we analyze the relation between markups (and profits) and overhead at the individual firm level, we find a strong positive relation, as expected. As we have pointed out all along, one of the reasons for raising prices and markups is that overhead has increased. The elasticity is 0.56 (see Table III): only just over half of the SG&A increases are passed on to markups. In a competitive economy this should be 1. Interestingly, firms with higher SG&A also have higher profits. In a competitive market, this coefficient should be 0. We can decompose the change in SG&A into R&D expenditure and advertising expenditure. These are often signaled as the components of SG&A that are important for intangible capital. Indeed, R&D expenditure has risen from 5% in 1980 to 20% of SG&A, and advertising from 4% to 10%. Even in 2016, these remain relatively minor shares of SG&A. The majority is still sales related and administrative expenditure. We find that the elasticity of R&D expenditure on markups is 16% and 5% for advertising expenditure. Interestingly, most of that effect remains when the dependent variable is the profit rate. This elasticity should be 0 under competition. Most of R&D and advertising

TABLE III

REGRESSIONS: EFFECT OF SG&A, R&D EXPENDITURE, AND ADVERTISING EXPENDITURE ON MARKUPS AND PROFIT RATE; EXTENSIVE MARGIN EFFECT OF R&D AND ADVERTISING

	Markup (log)			Profit rate (log)	
	(1)	(2)	(3)	(4)	(5)
SG&A (log)	0.56			0.15	
	(0.01)			(0.03)	
R&D exp. (log)		0.16			0.10
		(0.01)			(0.01)
Advertising exp. (log)		0.05			0.03
		(0.00)			(0.01)
R&D dummy			0.06		
			(0.01)		
Advertising dummy			−0.00		
			(0.03)		
R^2	0.61	0.07	0.43	0.04	0.05
N	26,743		247,615	26,743	

expenditures translate into profits as much as they do into higher markups. These are all at the intensive margin. When we evaluate the extensive margin—whether a firm does or does not have expenditures on R&D or advertising—we find an elasticity of 6% from R&D and no significant effect from advertising (since nearly all firms have advertising expenditure, there is not enough variation; only about 10% of the firms report R&D expenditure).

In sum, at the firm level, we find consistent evidence that profits and the market valuation of firms have gone up together with markups. Markups are not higher only to compensate for higher fixed costs, they are also higher because firms exert market power.

IV.B. Aggregate Profits and Markups

Even though markups and profit rates are different concepts—most notably because of the inclusion in profits of total costs, including overhead costs—they are related. In particular, there is an identity that links profit rates and markups and that holds for any technology $C(Q)$, as has been pointed out by Syverson (2019) and De Loecker and Eeckhout (2018b):

$$\pi_{it} = \frac{P_{it}Q_{it} - C(Q_{it})}{P_{it}Q_{it}} = 1 - \frac{AC_{it}}{\mu_{it}MC_{it}}, \tag{15}$$

where $\frac{AC_{it}}{MC_{it}}$ is the ratio of average cost to marginal cost and because $AC_{it} = \frac{C(Q_{it})}{Q_{it}}$ and $\mu_{it} = \frac{P_{it}}{MC_{it}}$.

Now there is a puzzle. The aggregate markup of 1.61 that we calculate in 2016 cannot be reconciled with the profit rate of 8%. In particular, Basu (2019) has pointed out that something must be wrong with our markup measure, because the implied profit is too high. If we plug in the aggregate markup in 2016 and assume that the ratio of average cost to marginal cost is equal to 1, then the implied profit rate is 38%.[42]

There are two problems with this argument. The first is that in this thought experiment, we have assumed that the average to marginal cost ratio is constant and equal to 1. We know from Figure VII that the fixed cost is sizable and has gone up. Therefore the average to marginal cost ratio is neither constant nor equal to 1.

The second problem with this argument is that it erroneously relies on a representative-firm framework. Equation (15) strictly holds at the firm level. In the aggregate, this translates into:

$$\pi_t = \sum_i m_{it}\pi_{it} = 1 - \sum_i m_{it}\frac{AC_{it}}{\mu_{it}MC_{it}} \neq 1 - \frac{AC_t}{\mu_t MC_t}, \tag{16}$$

where $\mu_t = \sum_i m_{it}\mu_{it}$, $AC_t = \sum_i m_{it}AC_{it}$, $MC_t = \sum_i m_{it}MC_{it}$. Therefore, the premise of a representative-firm framework is counterfactual.

Once we correct for these counterfactual assumptions—that the average cost to marginal cost ratio has increased and that we properly aggregate without assuming a representative-agent framework—the implied average profit rate of 8% in 2016 and the markup of 1.61 are indeed consistent.

In Figure XI, we decompose equation (15). We also report the actual values in Table IV.[43] When we assume both a

42. Basu (2019) performs a slightly different exercise. He rewrites equation (15) as $\mu_{it} = \frac{1}{1-\pi_{it}}\frac{AC_{it}}{MC_{it}}$ and takes the ratio between this expression evaluated at any two years, say, 2016 and 1980: $\frac{\mu_{2016}}{\mu_{1980}} = \frac{1-\pi_{1980}}{1-\pi_{2016}}$. He lets $\pi_{1980} = 0$ (which is close to the profit rate of 1% we find), then $\frac{1.61}{1.21} = \frac{1}{1-\pi_{2016}} \Rightarrow \pi_{2016} = 25\%$. This leads to a profit rate of 25%. This is completely unrealistic, especially since in our sample we find a profit rate of around 8% in 2016.

43. We use the technological specification of our benchmark model with Cobb-Douglas production and a fixed cost.

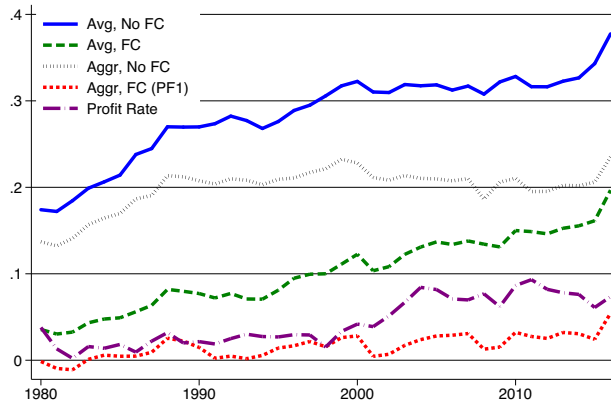


FIGURE XI

Decomposition of Equation (15) due to Overhead Costs and Aggregation

TABLE IV

DECOMPOSITION OF THE AVERAGE PROFIT RATE FROM EQUATION (15)

		Profit rate	
		1980	2016
Use average	No fixed cost	0.17	0.38
Use average	Fixed cost	0.04	0.20
Aggregation	No fixed cost	0.14	0.24
Aggregation	Fixed cost	0.00	0.05

representative firm and a constant average to marginal cost ratio equal to 1 (no fixed cost), we see profits rise from 18% in 1980 to 38% in 2016 (solid line). When we adjust for the observed average to marginal cost ratio but keep the representative firm assumption (long dashes in print; solid green in the color version online), the profits drop by more than half over the entire period. When we adjust for proper aggregation (drop the representative firm) and keep a constant average to marginal cost ratio (very short dashes in print; solid black in the color version online), profits drop by about one-third. Note that the gap is larger towards 2016 than in 1980, which is consistent with the fact that the distribution of firm sizes and markups has become more dispersed, resulting in a bigger gap between the aggregate and the average (due to Jensen's inequality). Finally, when we adjust for both proper aggregation and the observed

average to marginal cost ratio (dash-dot in print; solid purple in the color version online), profits are close to the observed profits in the data (short dashes in print; solid red in the color version online).

Overall, the relation that predicts average profit rates as a function of markups fits the data once we properly account for returns to scale (fixed costs) and once we properly aggregate. This indicates that our measure of markups does not predict an outlandish profit rate. What it does confirm is that markups and profit rates are different objects and that we should be careful comparing them. Too often, they are used interchangeably.

Finally, Traina (2018) proposes a different measure of market power that includes both COGS and SG&A. His measure is therefore closely related to the profit rate. Denote by τ:

$$\tau_{it} = \theta^{V+X} \frac{S_{it}}{P_{it}^V V_{it} + P_{it}^X X_{it}}, \tag{17}$$

where $p^V V$ is the expenditure on the variable input, $p^X X$ is the expenditure on overhead as measured by SG&A and where $\theta^{V+X} \approx 0.95$ (though he estimates a separate elasticity for each sector).

This ratio is directly related to the operating profit rate, the definition of which is

$$\pi_{it}^{OPX} = \frac{S_{it} - P_{it}^V V_{it} - P_{it}^X X_{it}}{S_{it}} = 1 - \frac{P_{it}^V V_{it} + P_{it}^X X_{it}}{S_{it}}. \tag{18}$$

We can therefore write the measure τ_{it} as

$$\tau_{it} = \theta^{V+X} \frac{1}{1 - \pi_{it}^{OPX}}. \tag{19}$$

Given this identity, this measure is closely related to the aggregate operating profit rate $\pi_t^{OPX} = \sum_i m_{it} \pi_{it}^{OPX}$ (see Figure F.1 in Appendix F). We find an increase in the operating profit rate between 1980 and 2016 of about 7–8 percentage points, and for the measure τ, which we interpret as an alternative measure of the profit rate, we see an increase of about 10 points, from 1.08 to 1.18.

In sum, aggregate markups and profitability are both increasing. Therefore the rise in markups is not exclusively due to the rise in overhead costs. This is evidence of the rise in market power.

V. The Macroeconomic Implications

The focus of our analysis so far has been on documenting in detail the time-series and cross-sectional evolution of markups and profitability. We now turn to discussing the macroeconomic implications of the rise in market power in the past decades.

V.A. The Secular Decline in the Labor Share

In the national accounts, the labor share of income measures the expenditure on labor (the wage bill) divided by the total income generated (value added). Although there are business cycle fluctuations, the labor share has been remarkably constant since World War II up to the 1980s, at around 62%. Since 1980, there has been a secular decline all the way down to 56% (Bureau of Labor Statistics Headline measure).[44] The decline since the 1980s occurs in the large majority of industries and across countries (see Gollin 2002; Karabarbounis and Neiman 2013).

Economists have struggled to understand the mechanism behind the decline in the labor share. One obvious hypothesis, ex ante, would be a within-firm substitution of labor for capital. This hypothesis is explored most prominently in Karabarbounis and Neiman (2013), which argues that a secular decrease in the relative price of investments goods led firms to substitute away from labor toward capital and can explain half of the decline in labor's share of income. The basic problem with this mechanism is that it rests crucially on a high elasticity of substitution between capital and labor (higher than 1). While Karabarbounis and Neiman (2013) claim that this elasticity is 1.25, the overwhelming majority of several decades of empirical studies (Antràs (2004), among many others) find that this elasticity is much lower than 1. The combination of a low elasticity of substitution between capital and labor, with the fact of a declining labor share of income, has been especially puzzling.

Koh, Santaeulalia-Llopis, and Zheng (2017) offer yet another explanation, which is based on the increasing importance of intangible capital and its incomplete measurement as part of capital in aggregate data. Firms now invest substantially more in intellectual property products, and this leads to a lower expenditure

44. There are issues of measurement. See Elsby, Hobijn, and Şahin (2013) on the role of how labor income of the self-employed is imputed. Even after adjusting for measurement issues, the labor share still exhibits a secular decline.

on labor.[45] However, in their world with perfect competition, this measurement issue should not lead to an increase in the total profit share. As we have documented, there is a substantial increase in the profit rate. If intangibles play a role, it must allow firms to exert more market power, which is the central thesis of our article. We do find evidence that expenditure on overhead has increased (see below), which could certainly include intangibles, but we also find that economic profits increase even if we interpret overhead (and hence intangibles) as a factor of production (see Figure VIII, Panel A). Finally, Elsby, Hobijn, and Şahin (2013) find little support for capital-labor substitution, nor for the role of a decline in unionization. They do find some support for offshoring labor-intensive work as a potential explanation.

In the context of our setup, the change in the markup has an immediate implication for the labor share. Although we have calculated the markup from all variable inputs, we could do so as well for labor alone. Then rewriting the first-order condition (7) where $V = L$, $P^V = w$, and $\theta^V = \theta^L$, the output elasticity of labor, we obtain that at the firm level the labor share satisfies

$$\frac{w_t L_{it}}{P_t Q_{it}} = \frac{\theta_{it}^L}{\mu_{it}}. \tag{20}$$

Observe that if there are multiple inputs that are fully variable, the estimated markup should be the same. So even if the markup is calculated for the bundle V, it should also hold for L as long as both V and L are variable. Profit maximization by individual firms thus implies that the labor share is inversely proportional to the markup. As the markup increases, we expect to see a decrease in the labor share.

Unfortunately Compustat does not have good data for the wage bill. Because reporting compensation to the SEC is not compulsory, the variable XLR for total compensation is heavily underreported.[46] Because of selection in the sample of those firms that do report total compensation, we need to be cautious interpreting the aggregate labor share outcomes.

45. Intangible assets are nonphysical assets including patents, trademarks, copyrights, and franchises, that grant rights and privileges and have value for the owner.

46. Less than 10% of the year-firm observations include XLR.

TABLE V
REGRESSIONS: LOG (LABOR SHARE) ON LOG (MARKUP)

	Labor share (log)					
	(1)	(2)	(3)	(4)	(5)	(6)
Markup (log)	−0.24	−0.23	−0.20	−0.24	−0.68	−0.73
	(0.03)	(0.03)	(0.03)	(0.03)	(0.02)	(0.02)
Cost share (log)					0.91	0.96
					(0.01)	(0.01)
Year FE		X	X	X	X	X
Industry FE			X		X	
Firm FE				X		X
R^2	0.02	0.08	0.21	0.88	0.93	0.99
N	24,838					

Note: FE = fixed effects. Four-digit industries. Standard errors (in parentheses) are clustered at the firm level.

Despite the shortcomings of our data, we can nonetheless verify the firm's optimization condition (20) at the firm level. In Table V we report the regression coefficients of the log of the labor share on the log of the firm's markup. The first four specifications only differ in the fixed effects that are included. We consistently find a negative coefficient of around −0.20 to 0.24. As a firm's markup increases by, say, 10%, its labor share decreases by 2–2.4%.

To extrapolate these firm-level results to the aggregate economy, we need to keep in mind that there is no such thing as a representative firm in this context. The rise of average markups is distributed unequally, and increasingly so. Most important, since two-thirds of the rise in market power is due to reallocation of economic activity toward high-markup firms, the effect of markups on the labor share in the aggregate is predominantly driven by a few large firms with high markups and a low labor share. Our findings for the firm-level markups are thus consistent with those in Autor et al. (2020) and Kehrig and Vincent (2017) for the Census of Manufacturing. In sum, we find firm-level evidence of the direct inverse relation between markups and the labor share that we obtain from the first-order condition (20).

In the table, we also analyze whether we can reject any evidence that there is perfect competition. The fifth and sixth columns report the same regression where we now include the log of the cost share (labor over total cost) as a covariate. Under perfect competition, the coefficient is 1. Here we find a coefficient

significantly smaller than 1, indicating that there is a wedge between sales and costs. Equally important, any other covariate (in this case markup) should be insignificant. We find instead that the coefficient on the markup is highly significant and negative. This indicates that there is evidence of noncompetitive price setting.

V.B. The Secular Decline in the Capital Share

The same logic for the decline in the labor share also applies to materials M, that is, variable inputs that are used in production. Those are included in our variable cost measure COGS. Now if we consider the evolution of capital expenses, which is not included in our measure of variable cost and which adjusts at a lower and more long-run frequency, then the increase in markup has implications for the capital share.[47] In the long run and once the adjustment frictions are taken into account, higher output prices and lower output quantities eventually will lead to a decrease in the capital share. While the decline in the labor share is widely discussed, the decline in the capital share has received much less attention.[48]

Assuming a static environment, the following equality has to hold:

$$\frac{P^V V}{PQ} + \frac{rK}{PQ} = 1 - \frac{P^X X}{PQ} - \frac{\Pi}{PQ}, \tag{21}$$

The labor share and the capital share sum up to 1 minus the profit share minus the overhead share. We have established that the profit share and the overhead share increase, so the right-hand side decreases. With complementary capital and variable inputs, and over a long enough time horizon for capital to adjust, the expenditure on capital rK as a share of output will be decreasing over time. In fact, if capital were fully flexible, it would adjust

47. This is independent of the frequency at which capital adjusts. Implicit in our assumptions is the fact that variable inputs, which consist of labor L and material inputs M, fully adjust within a year, our unit of time. This assumption allows us to calculate the markup. Capital may or may not adjust. From Online Appendix Figure 1.3a, we have inferred that capital is not equally flexible as the variable inputs.

48. A notable exception is Barkai (2017). He uses aggregate data: value added and compensation from the National Income and Productivity Accounts, and capital from the Bureau of Economic Analysis Fixed Asset Table. Instead, we use firm-level data.

TABLE VI
REGRESSIONS: LOG(CAPITAL SHARE) ON LOG(MARKUP)

	Capital share (log)					
	(1)	(2)	(3)	(4)	(5)	(6)
Markup (log)	0.03	0.03	−0.02	−0.14	−0.90	−0.86
	(0.02)	(0.02)	(0.01)	(0.02)	(0.00)	(0.00)
Cost Share (log)					1.13	1.11
					(0.00)	(0.00)
Year FE		X	X	X	X	X
Industry FE			X		X	
Firm FE				X		X
R^2	0.00	0.02	0.31	0.83	0.98	1.00
N	242,692					

Note: FE = fixed effects. Four-digit industries. Standard errors (in parentheses) are clustered at the firm level.

according to the equivalent of first-order condition (7) $\frac{rK}{PQ} = \frac{\theta^K}{\mu}$ which relates the capital share to the inverse of the markup.

In Online Appendix Figure 14.1b we document the evolution of the capital share for the firms in our data. Not surprisingly this measure is quite volatile because it is a long-term measure that adjusts at a lower frequency and is more subject to aggregate fluctuations. Also, before the 1980s, capital investment was particularly low because of tumultuous financial times: inflation was high and financial frictions were considered higher. What we learn from the figure is that there was a decrease in the capital share from around 12% in 1980 to 8–10% toward the end of the sample. In the aggregate, the capital share is correlated with the inverse of our markup measure. With a long enough horizon, capital investment adjusts and hence there will be a reduction in capital investment as markups increase.[49]

As with the labor share, we can also investigate the firm-level relation between the capital share and markups. In Table VI we report the regression coefficients for different specifications. We find that without firm fixed effects, there is no significant relation between markups and the capital share. This may be indicative

49. A more detailed analysis of the impact of market concentration on business investment is in Gutiérrez and Philippon (2017). In particular, they show within manufacturing that there is a positive investment response to competition from China.

of the adjustment costs that firms face when investing in capital. Instead, with firm fixed effects, there is a significant negative effect, with an elasticity of −0.14. When we include the cost share, the coefficient on the cost share is larger than 1. Under variable adjustment of capital, perfect competition would require this to be equal to 1, and less than 1 with market power (see, for example, Table V for the labor share). The fact that the coefficients on the cost share here are larger than 1 indicates that capital does not adjust frictionlessly.

V.C. The Secular Decline in Low-Skill Wages and Labor Force Participation

An increase in markups implies a decrease in aggregate output produced, whenever demand is not perfectly inelastic. Lower output produced then implies lower demand for labor. This results in both lower labor force participation and lower wages. Even if supply is perfectly elastic, real wages decrease with market power because the price of the output goods has increased.

There is ample evidence of the stagnation of wages in the lower half of the distribution. The median weekly wage in constant prices has changed barely since 1980, from $330 to $345 (1982 prices, source Current Population Survey). But there has been technological progress, and the share of median wages out of GDP has nearly halved, because over three and a half decades GDP has nearly doubled. In the past few decades, labor force participation has also been decreasing from 67% in the 1990s to 63% now. Most strikingly, while the gender gap has continued to close, in the past two decades female labor force participation is also decreasing.

The quantitative investigation of the effect of market power on low-skill wages and labor force participation is beyond the scope of the current article. In De Loecker, Eeckhout, and Mongey (2018) we construct an oligopolistic framework for firm dynamics that quantitatively accounts for these general equilibrium implications of the rise in market power. We find that market power indeed has an effect on equilibrium wages, and that quantitatively, that effect is large. Our quantitative model predicts that real wages as a share of GDP drop by over 26%, consistent with what we see in the data.

V.D. *The Secular Decline in Labor Reallocation and Migration Rates*

It is well known that in an environment with market power, shocks to productivity and costs are not translated one for one into prices. In a competitive market, firms face a perfectly elastic demand and any decrease in costs is passed on to the consumer, where prices decrease by the same amount as the decrease in costs. With market power however, the pass-through of cost shocks to prices is generally incomplete.[50] Crucial for our finding is that the higher the degree of market power by firms, the lower the pass-through.

Now consider an environment where firms have market power and face shocks to their productivity. With positive shocks, firms face lower costs and adjust their inputs (say, labor) upward. With negative shocks, they adjust inputs downward. Because pass-through is lower in the presence of higher market power, the rise in market power will give rise to lower degree of adjustment of the variable inputs, including labor, for the same shock process.

This is precisely what Decker et al. (2014) find for the U.S. economy over the past three decades. The volatility of shocks has not decreased, but rather the responsiveness of firm's output and labor force decisions to the existing shocks has declined.[51] The rise in market power can thus rationalize the decrease in labor reallocation across firms, even if the observed shocks to firm productivity have remained constant.

The decrease in labor market dynamism is evident in the decrease of labor reallocation as well as in the decrease of job-to-job transitions, nonemployment to employment transitions, and

50. Most of the evidence comes from studies that measure the impact of changes in the exchange rate or reductions in tariffs; see, for example, Campa and Goldberg (2005). More recently, incomplete pass-through has been documented in a domestic setting. For example, Ganapati, Shapiro, and Walker (2018) reports incomplete pass-through of energy input price changes across industries of the U.S. manufacturing sector.

51. Independent evidence at business cycle frequency by Berger and Vavra (2017) establishes that the volatility of prices is due to firms' time-varying responsiveness to shocks rather than to the time-varying nature of the shocks themselves. Their identification strategy is derived from the exchange rate pass-through of volatility on prices.

employment to nonemployment transitions.[52] The decrease in market power and the resulting decrease in labor reallocation can also rationalize the fact that migration rates across U.S. states and metropolitan areas have decreased by nearly half from around 3% in 1980 to 1.5% in 2016.[53] If firms are based in different local labor markets and a fraction of all job relocation decisions are between local labor markets, then lower job flow rates will automatically give rise to lower migration rates. We assess the quantitative significance of the impact of the rise of market power on labor reallocation and migration in a companion work (De Loecker, Eeckhout, and Mongey 2018).[54]

VI. Discussion and Robustness

Here we discuss the features of our model and report a number of robustness exercises.

VI.A. Cost Shares

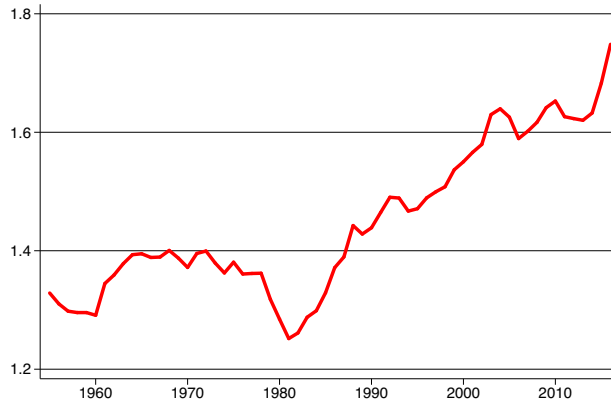
We repeat the analysis where we obtain the output elasticity from cost shares. For each firm, we have an observation for the cost share $\alpha_{it}^{V} = \frac{P_t^V V_{it}}{P_t^V V_{it} + r_t K_{it}}$. Within an industry, we use the median of the distribution as the measure for the output elasticity: $\theta_{st} = \text{median}_{i \in s}\{\alpha_{it}^{V}\}$.

Figure XII, Panel A reports the sales-weighted average of the markups with the output elasticity derived from the cost share for the traditional production technology where overhead is a fixed cost and denoted by CS. The pattern is very similar to that in Figure I. There is a moderate decrease from the 1960s and then

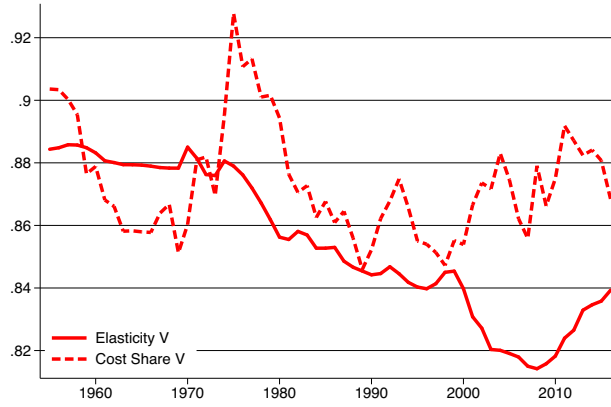
52. There are several potential alternative explanations for the decline in job flows: demographic change (aging workforce; Fallick, Fleischman, and Pingle 2010; Engbom 2017), a more-skilled workforce, lower population growth, decreased labor supply (Karahan, Pugsley, and Şahin 2016), technological change (Eeckhout and Weng 2017), changed volatility of production, and government policy (such as employment protection legislation, or licensing; see Davis and Haltiwanger 2014). Hyatt and Spletzer (2013) show that demographic changes can explain at most one-third of the decline in job flows.

53. See Kaplan and Schulhofer-Wohl (2012), among others.

54. Recent work by Baqaee and Farhi (2019) also draws attention to the fact that firm-level productivity shocks can give rise to a nonlinear impact on macroeconomic outcomes. For example, models with network linkages such as Gabaix (2011) give rise to such nonlinearities. The framework in De Loecker, Eeckhout, and Mongey (2018) establishes that market power in the presence of incomplete pass-through also gives rise to nonlinearities.



(A) Aggregate markup (CS)



(B) Output elasticity and cost shares: V

FIGURE XII

Cost-Share Based Aggregate Markups and Technology

Panel A reports the aggregate markups using cost shares (median cost shares for each year and sector). Panel B compares the sector-weighted cost share to the estimated output elasticities; weights are sectoral total sales.

an increase from 1980 up to 2016. The level is slightly higher, and the increase by 50 points is somewhat more pronounced.

From inspection of the definition of the markup in equation (7), the rise in the markup could potentially be attributed to two sources: (i) an increase in the ratio of sales to expenditure

on variable inputs; or (ii) technological change, an increase in the output elasticity θ^V over time. In Figure XII, Panel B we plot the average cost share of the factors of production V and K as well as the average output elasticity estimated from the benchmark technology. There is some volatility in the cost shares, but they are in line with the estimated output elasticity. This indicates that the steep increase in markups is driven by the increase of sales over expenditure on inputs. Firms are selling their goods at higher margins. This is also evident from inspection of Figure II, Panel B, confirming again that the evolution of the share-weighted average markup is mainly driven by the ratio of sales to expenditure on variable inputs and not by changes in the output elasticity.

VI.B. Production Function with Overhead as a Factor of Production

The conventional production function uses as factors of production the variable input V and capital K. All other expenditures accounted for as not directly related to the production of the goods sold are overhead. They are considered fixed costs, a cost incurred that is independent of the output produced. This is the standard approach in the industrial organization literature on markup estimation.

In contrast to the conventional interpretation of the production technology, we propose an alternative interpretation where a portion of the overhead is a factor of production. Higher expenditure on getting more and better logistics managers will lead to an increase in the units produced. More sales people increases the units sold. To interpret overhead as a factor of production, we denote its expenditure by $p^X X$, where the quantity that enters the production technology is X and the unit price is p^X.

We now take this nonconventional interpretation of overhead as a factor of production seriously and assume that all of it is a factor. The production function can then be written as $Q(V, K, X)$ and firm profits are $PQ(V, K, X) - P^V V - rK - P^X X$. We can apply the same cost-based method for the derivation of markups as laid out in Section II. We treat X as a factor of production that enters the production function, but it is nonvariable, just like capital K. The treatment of the variable input V remains as before. The difference relevant for the measurement of markups therefore stems from the production function estimation and the resulting estimate for θ^V. To differentiate, we denote the estimates from

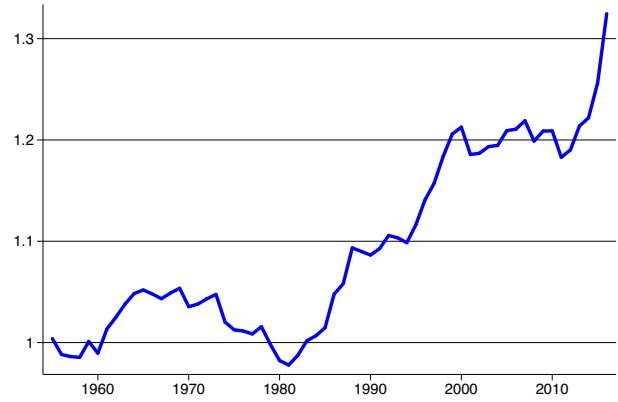
this production function by PF2. When we calculate elasticities based on cost shares that take into account overhead as a factor of production, we refer to it as CS2. To further differentiate the graphical representation, PF1 and CS1 are plotted in red, and PF2 and CS2 are in blue (color versions of all figures are available online).

Figure XIII, Panel B plots the cost share of variable factors in the total cost (consisting of variable factors, capital, and overhead), as well as the cost share of overhead. We see that there is a slight decrease in the cost share of the variable factor of production from 80% in the beginning of the sample to 70% in 2016. The share of the fixed cost has increased from 18% at the beginning to 24% toward the end. This is indicative of the fact that the overhead cost, and thus the technology, has changed. The estimated output elasticities confirm this pattern, although importantly they do not necessarily have to sum to one (including rK of course).

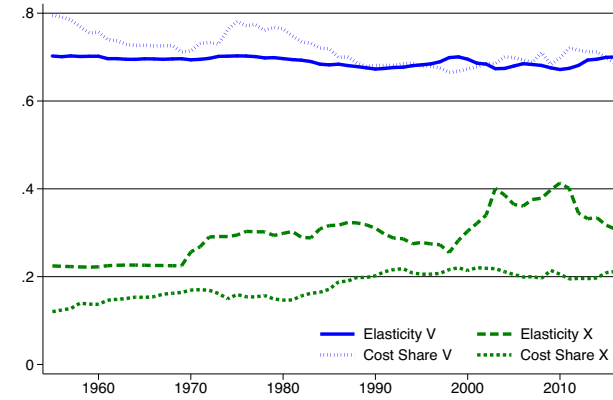
In Figure XIII, Panel A, we report the evolution of the average markup with this new production technology. Qualitatively, we see a similar pattern for the increase in the average markup starting in 1980. Initially around 1, the average markup increases by about 30 points by 2016. The increase for this technology is 10 percentage points lower than for the traditional production function (Figure I). This difference is driven by the fact that the cost share of overhead (and the estimated output elasticity θ^X) is increasing over time (see Figure XIII, Panel B). What matters for the markup estimate however is the elasticity θ^V. We know that it is roughly constant for the conventional production function (Figure XII, Panel B). For the production technology with overhead as a factor of production, θ^V is slightly decreasing (Figure XIII, Panel B). Therefore the estimated markup shows a more moderate increase (30 points) than under the conventional production technology (40 points).

VI.C. Returns to Scale

With the estimated technologies, we can evaluate any technological change that affects the returns to scale. Because the technology is Cobb-Douglas, the returns to scale are measured by the sum of the output elasticities: $\theta^V + \theta^K$ for PF1 and $\theta^V + \theta^K + \theta^X$ for PF2. We find that the estimated technology shows a rise in the degree of increasing returns over time. In Figure XIV, Panel A we report the sum of the output elasticities



(A) Markup with θ_{st} from PF2



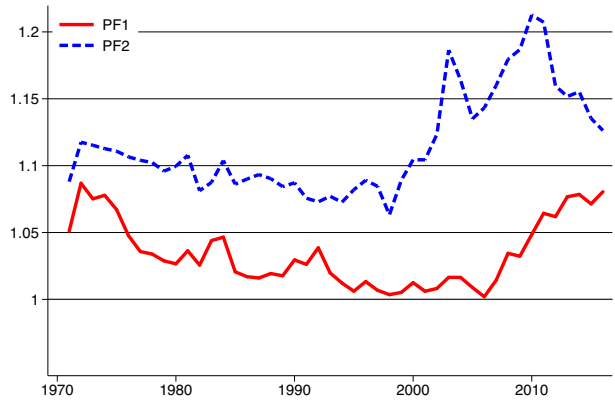
(B) Output elasticity PF2 and cost shares: V and X

FIGURE XIII

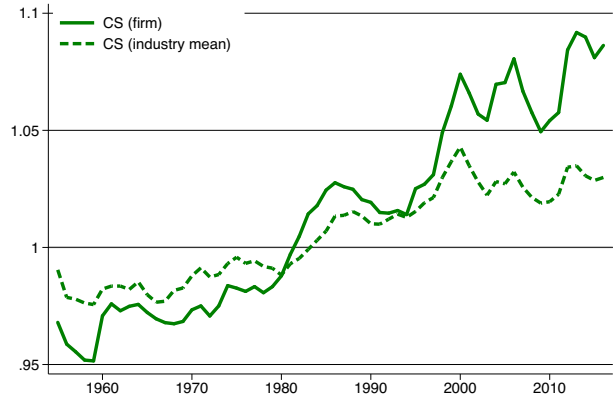
Average Markups, Elasticities, and Cost Shares for Production Function with Overhead as a Factor

Output elasticities from estimated PF2 and from CS2: time-varying, sector-specific (two-digit) output elasticity θ_{st} (revenue-weighted average).

for both technologies PF1 and PF2. For the conventional technology (PF1), from the start of the sample, the estimated returns to scale go from around 1.02 in 1980 to 1.08 in 2016. We estimate an increase in the returns to scale of the technology with overhead as a factor of production, from 1.07 up to 1.13, reaching 1.22 in 2010. The fact that the production function with overhead as



(A) Returns to scale (sum of output elasticities) of estimated PF1 and PF2; revenue weighted



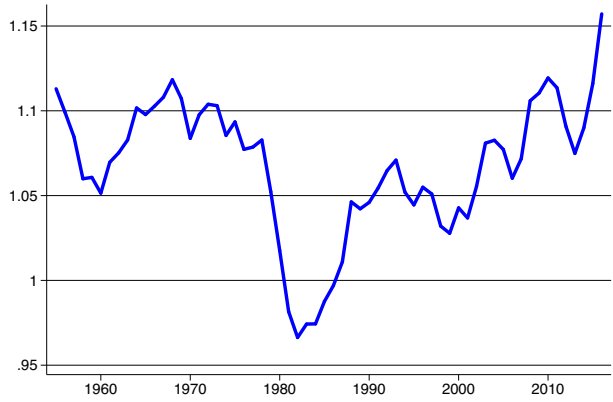
(B) Estimated returns to scale of cost shares: firm CS and sector average CS

FIGURE XIV

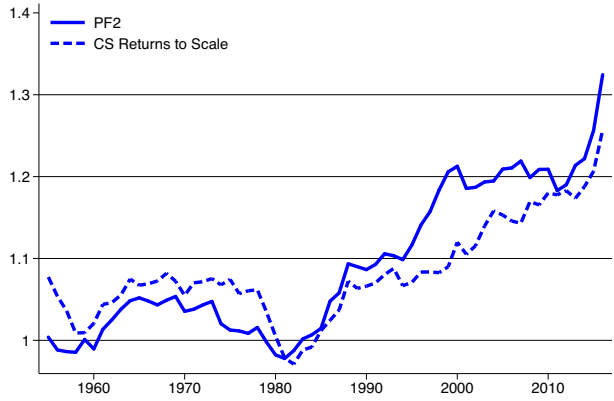
Returns to Scale

an input has higher returns to scale confirms that overhead X is in part a fixed cost that generates increasing returns. Moreover, those returns to scale are increasing more over time as overhead increases, which establishes that the role of overhead as a source of returns to scale is growing.

An alternative way to measure returns to scale is with a method first used in Syverson (2004). While using cost shares



(A) Markup with θ_{st} from CS2



(B) Average markup: benchmark (PF2) and cost shares with returns to scale (Syverson)

FIGURE XV

Average Markups for Production Function with Overhead as a Factor

Output elasticities from estimated PF2 and from CS2: time-varying, sector-specific (two-digit) output elasticity θ_{st} (revenue-weighted average).

implicitly assumes that the technology is constant returns, Syverson (2004) adjusts the technology based on cost shares and derives the returns to scale. He assumes the following functional form for the technology based on cost shares but without constant

returns:

$$q = \gamma \left[\alpha_V v + \alpha_K k + \alpha_X x\right] + \omega, \tag{22}$$

with all variables in logs, where $\alpha_V = \frac{P^V V}{P^V V + rK + P^X X}$ is the cost share of the variable input, and likewise for α_K and α_X.

While each cost share determines the output elasticity, the technology need not be constant returns and the curvature is captured by γ. In Figure XIV, Panel B we plot two measures of the estimated γ, one for the average firm-level γ and one where we impose a common γ at the two-digit industry level for the technology with overhead as a factor of production.[55] These graphs reveal that also with this method, returns to scale have increased throughout the sample. There were decreasing returns to scale before 1980 and since 1980 returns to scale have been increasing, up to 1.05 at the end of the sample.

The increase in the returns to scale also explains why the markup estimate based on cost shares only shows an increase of 20 percentage points (Figure XV, Panel A), whereas under the elasticity estimated from the production function the increase since 1980 is 30 percentage points (Figure XIII, Panel A). By construction, cost shares add up to 1, and therefore the implied elasticities are derived under the assumption of constant returns. As a result, the increase in the elasticity θ^X due to an increase in the expenditure share of overhead must necessarily lead to a decrease in θ^V.[56] With θ^V decreasing, from equation (7), the increase in the markup must necessarily be dampened. This illustrates that directly using the cost shares can by construction not account for any change in the returns to scale in the technology.

The evolution of returns to scale helps us understand the difference between Figures XIII, Panel A and XV, Panel A. In the latter, we ignore the change in the returns to scale because cost shares are implicitly assuming CRS. If instead we use the elasticities obtained for the Syverson (2004) technology in equation (22), which is equal to $\gamma\alpha_v$, we obtain an average markup (see Figure XV, Panel B) that is very similar to the one using

55. An important caveat here is that we estimate this technology by means of a simple regression, without accounting for endogeneity, because the cost share approach implicitly assumes all inputs adjust within the time period.

56. In principle, it could also lead to a decrease in θ^K, but θ^K is so small it cannot offset all of the increase in θ^X. We consistently find that the estimated θ^K is constant across all specifications.

TABLE VII

AGGREGATE MARKUPS: VARIATION BY TECHNOLOGY (θ) AND WEIGHTING (m_{it})

Output elasticity	Revenue weight	Input weight		
		1 Input		All inputs
	$m_{it} = \frac{R_{it}}{R_t}$	$m_{it} = \frac{COGS_{it}}{COGS_t}$	$m_{it} = \frac{L_{it}}{L_t}$	$m_{it} = \frac{TC_{it}}{TC_t}$
Economy-wide θ	$\theta \sum_i m_{it} \frac{R_{it}}{COGS_{it}}$	$\theta \frac{R_t}{COGS_t}$	$\theta \sum_i m_{it} \frac{R_{it}}{COGS_{it}}$	$\theta \sum_i m_{it} \frac{R_{it}}{COGS_{it}}$
Sector-specific θ_{st}	$\sum_i \theta_{st} m_{it} \frac{R_{it}}{COGS_{it}}$	$\sum_s \theta_{st} \frac{R_{st}}{COGS_{st}}$	$\sum_i \theta_{st} m_{it} \frac{R_{it}}{COGS_{it}}$	$\sum_i \theta_{st} m_{it} \frac{R_{it}}{COGS_{it}}$

Note. $TC_{it} = COGS_{it} + r_t K_{it} + SGA_{it}$.

the elasticity estimated with the production function (PF2). The increase in γ in Figure XIV, Panel B implies that the elasticity α_V used in Figure XV, Panel A is multiplied by γ.

Finally, in Online Appendix 13 we also analyze the returns to scale using the data from the censuses.

VI.D. Input Weights and Joint Distributions

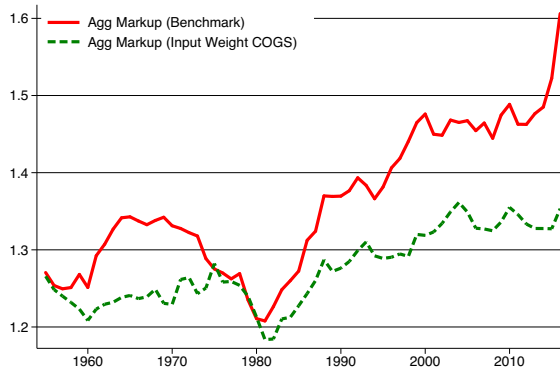
We have shown in Section III that to calculate aggregate markups, the choice of the weighting measure matters. Because we are interested in the entire joint distribution of markups and firm characteristics (revenue, costs, inputs, etc.), having information on as many moments as possible provides more detailed insights about the evolution of markups.[57]

Here we report the average markup using different weights. To get a better idea of the definition of each of the weights, we report them in Table VII.

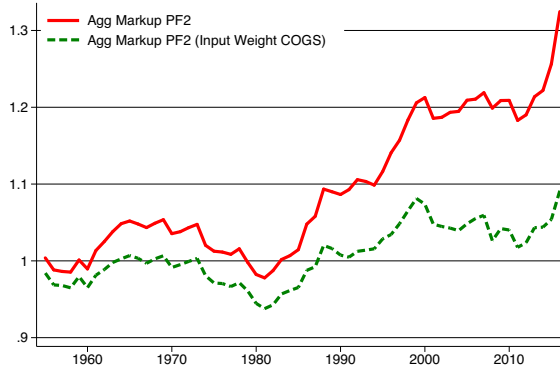
For our data, in Section III we have already reported the revenue-weighted aggregate markups in Figures I and II, Panel A, as well as the input-weighted markup with total cost as the input weight (Figure II, Panel A).

Here we plot the aggregate markups with one input weight and for a sector-specific output elasticity (the results are very similar with economy-wide, constant output elasticities). We present two versions of the production technology (PF1 and PF2). In row 1 (Figure XVI, Panels A and B) we plot the aggregate markup measure proposed by Edmond, Midrigan, and Xu (2015) and Grassi (2017) on the grounds of its representativeness of welfare

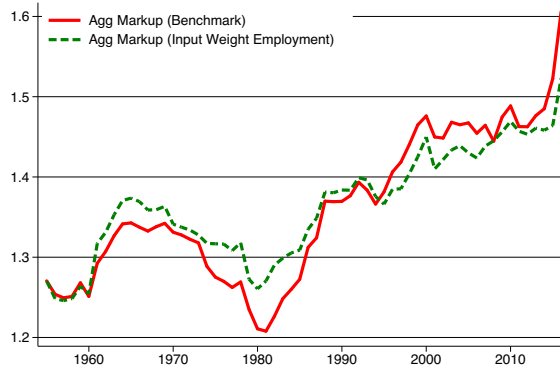
57. In Online Appendix 6 we report the contour plots of the joint distributions of markups and revenue and inputs. This gives us a view of the raw data.



(A) COGS, PF1



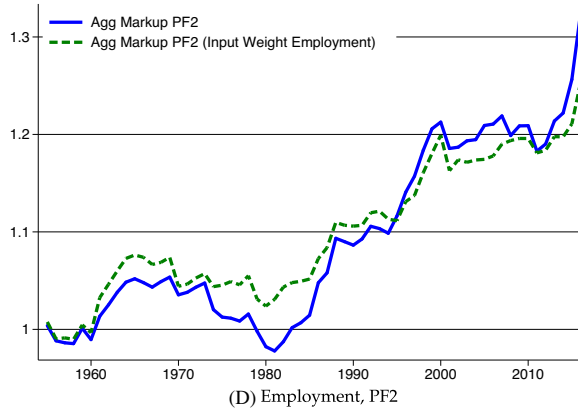
(B) COGS, PF2



(C) Employment, PF1

FIGURE XVI

Markups with Input Weights: COGS and Employment for the Benchmark Technology and PF2



(D) Employment, PF2

FIGURE XVI
(CONTINUED)

measures in a setting with CES preferences in models such as monopolistic competition and Atkeson and Burstein (2008). Here we do not take a stance on welfare, but we do in De Loecker, Eeckhout, and Mongey (2018).

We find a rise in the COGS-weighted aggregate markups that is only about half of the rise in revenue-weighted markups and is substantially lower than the total cost–weighted aggregate markup (Figure XVI, Panels A and B). We can see from Table VII why. With one input equal to the variable input used for calculating the markup, the aggregate markup is simply the ratio of revenue over COGS (multiplied by the output elasticity). As a result, the markup is a function of aggregates only. This implies that the aggregate is not sensitive to within-sector (or within-economy) variation. In fact, the aggregate markup measure is identical to that obtained by Hall (1988) and that we report in Figure V.

Rather than using expenditure, we can also use quantities. Although we do not have quantities of COGS (because we do not know the unit prices), we have quantities of labor, as measured by the number of employees. In Figure XVI, Panels C and D for the two technologies we use employment weights. Unlike aggregate markups with weights from expenditure shares of inputs, those with employment weights track the benchmark aggregates. This seems to indicate that at least for employment, the quantities do not adjust as much as the input prices (in this case wages)

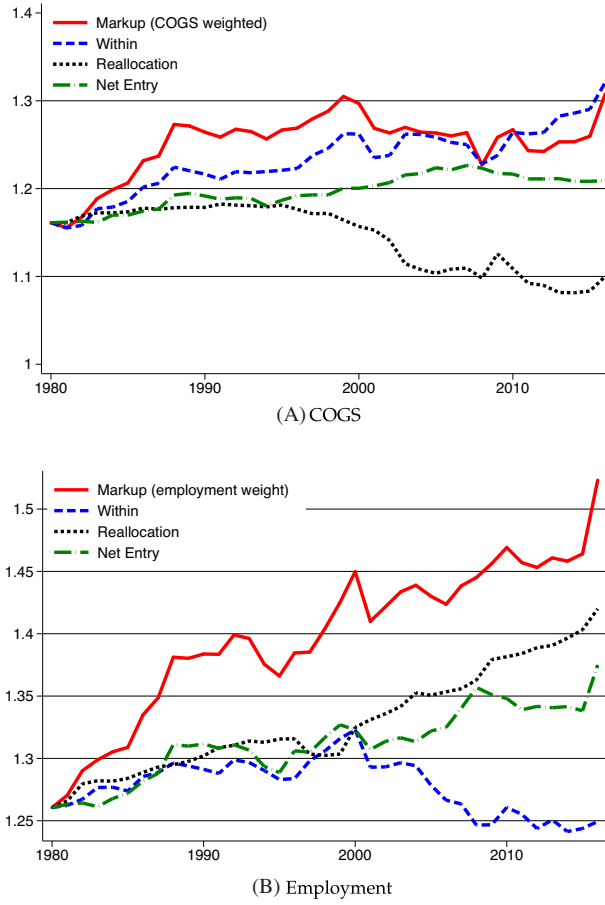


FIGURE XVII

Decomposition of Input-Weighted Average Markups

do. When we quantify the economy in De Loecker, Eeckhout, and Mongey (2018), we find large general equilibrium effects on wages and smaller effects on labor supply.

Finally, we also do the same decomposition exercise on our input-weighted measures in Figure XVII. Not surprisingly, for the COGS-weighted aggregate, there is no role for reallocation (the reallocation term is even negative). Because by construction the COGS-weighted measure is based on averages only, there is no impact of within-industry reallocation, and we have shown earlier

that the reallocation occurs predominantly within industry. What the decomposition shows is that virtually all of the change in the COGS-weighted markups is driven by the within term, the rise in markups themselves.

Instead, for the employment-weighted measure, the picture looks much more similar to that of the benchmark revenue-weighted aggregate markup. About two-thirds of the rise in the employment-weighted average markup can be attributed to reallocation.

An important conclusion to take away from these alternative measures for average markups is that they are different moments of a much richer distribution of markups. We have documented that the distribution has a fairly constant median, that the upper tail has become a lot fatter, and within a market, larger firms tend to have higher markups. The different weights give us further insights into the joint distribution of markups, revenue, and all inputs.

VI.E. Comparison of Our Estimates with Those in the Literature

In Online Appendix Section 7, we compare our estimates with those obtained in the literature using the demand approach (Berry, Levinsohn, and Pakes 1995) for seven industries for which there are data: beer, breakfast cereal, steel, autos, airlines, department stores, and electronic shopping and mail order. For the companies in our data set that fall in the same industry classification, we construct an average markup and plot them jointly with the markups obtained in the literature (Online Appendix Figure 7.1).

Whenever there is overlap, the patterns of markups obtained with the demand approach closely follow those obtained with our cost-based approach. This is remarkable because not only are the methods different, they rely on different data. This is testament to the fact that the estimates we obtain are robust across different methods and data sources.

We perform further robustness exercises in the Appendix and the Online Appendix.

VII. Concluding Remarks

Using firm-level data on the accounts of all publicly traded firms and of the census of private firms (in manufacturing, retail and wholesale trade) in the United States, we study the evolution

of market power. For each firm, we estimate both markups and profitability, and we document the properties of their distribution. We find that from 1980 onward, markups have risen from 21% to nearly 61% in 2014, an increase of 40 points. For the same period, average profit rates have increased from 1% of sales to 8%.

We attribute this rise in market power nearly exclusively to the increase for the firms with the highest markups already. The distribution of markups has become more skewed with a fat upper tail while the median of the distribution remains unchanged. Because of this increasingly skewed distribution, we must be cautious not to use the average markup as that of a representative firm to draw any conclusion about the aggregate economy. When markets are noncompetitive, aggregation is generally nonlinear. In particular, the rise in revenue-weighted markups is due in part to the rise of the markups themselves and in part to the reallocation of sales shares from low- to high-markup firms. We find that reallocation accounts for two-thirds of the rise.

We further establish that the rise in markups is not merely to offset a rise in overhead costs. Although overhead costs have risen, the rise in markups exceeds that of overhead. We thus find that there are excess markups, and that the excess markups are highest for those firms with high overhead costs. This is consistent with the increase in our measure of profits. We also find substantial increases in the market value as a share of sales. All this indicates that the rise in markups is evidence of a rise in market power.

We use our evidence to investigate the macroeconomic implications of the rise of markups. We focus our attention on the decrease in the labor share. From the first-order condition of the firm's optimization problem, there is a negative relation between the labor share and the markup. We establish that this negative relation exists at the firm level. This provides a compelling justification for the secular decline in the labor share that the aggregate U.S. economy has experienced. We further discuss the impact of the rise in market power on the decrease in the capital share, on the decrease in low-skill wages and labor force participation, and on the decrease in labor market dynamism and migration rates.

Markups of some firms are reaching heights multiple times higher than ever seen, at least since World War II, when our data start. It is open to speculation whether this trend will continue, but for now there are no signs that markups will decrease substantially any time soon.

APPENDIX A: ESTIMATING OUTPUT ELASTICITIES

A crucial component to measure markups is to obtain an estimate of the output elasticity of a variable input of production (θ^V). Although the production approach to markup estimation described in De Loecker and Warzynski (2012) does not restrict the output elasticity, when implementing this procedure, it depends on a specific production function and assumptions of underlying producer behavior to identify and estimate the elasticity in the data. We use two distinct methods to estimate the output elasticity of the production function. First, we estimate a parametric production function for each sector-year using recent techniques that take into account the well-known potential biases discussed in the literature. Second, we nonparametrically estimate the output elasticity using (constructed) cost shares. Both approaches have their advantages and disadvantages, which we discuss.

A.A. Production Function Estimation

We follow standard practice and rely on a panel of firms, for which we estimate production functions for each (two-digit) industry. For the benchmark specification, we consider a sector-year-specific Cobb-Douglas production function, with a variable input bundle and capital as inputs. For each industry s we consider

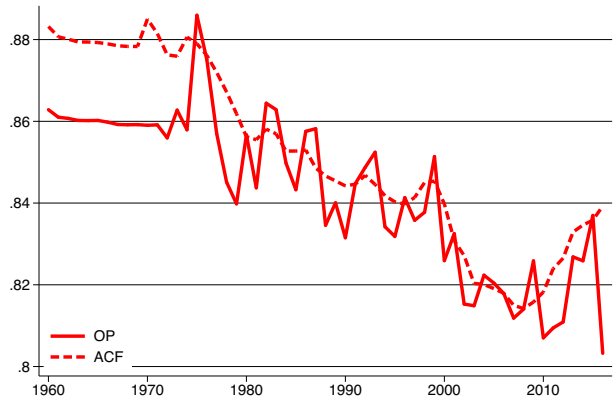


FIGURE A.1

Output Elasticities under Alternative Data-Generating Processes: Olley and Pakes (1996) and Ackerberg, Caves, and Frazer (2015)

the production function (PF1 in the main text):

$$y_{it} = \theta_t^V v_{it} + \theta_t^K k_{it} + \omega_{it} + \epsilon_{it}, \tag{23}$$

where lowercase letters denote logs and $\omega_{it} = \ln \Omega_{it}$, where y_{it} is a measure of realized firm's output, and ϵ_{it} captures measurement error in output—that is, $y_{it} = \ln(Q_{it}\exp(\epsilon_{it}))$.

We depart from the standard specification in the literature by considering time-varying production function parameters. In particular, in the baseline model we estimate production functions with both time-varying and sector-specific coefficients, for each of the 22 sectors (i.e., two-digit NAICS).[58] There are good reasons to believe that technology varies across sectors of the economy, from retail with giants like Walmart and Amazon, to highly specialized medical devices companies. Equally or more important for the evolution of markups is that the technology is time varying. Over a period of seven decades, technology is likely to change. This is important for the estimation of markups because systematic technological change will imply a time-varying output elasticity θ_{it}^V. From inspection of equation (7), imposing a constant technology and hence a constant θ^V will therefore yield an overestimate of the markup if θ^V is decreasing and an underestimate if θ^V is increasing.[59] Allowing the production function coefficients to vary over time is also a parsimonious way to account for factor-biased technological change.

When we consider the production function with overhead (PF2 in the main text), the specification is given by:

$$y_{it} = \theta_t^V v_{it} + \theta_t^K k_{it} + \omega_{it} + \theta_t^X x_{it} + \epsilon_{it}, \tag{24}$$

with $x = ln(X)$, and X captures (deflated) SG&A.

The challenges in estimating production functions, using any data set, be it the Compustat data or plant-level manufacturing census data, can be grouped into two main categories: dealing with unobserved productivity shocks (ω_{it}); and extracting units of output and inputs from revenue and expenditure data (i.e., the

58. In principle we can consider industries at a lower level of aggregation, for example, three-digit NAICS, at the cost of pooling over longer periods of time, hereby keeping output elasticities constant over that same period.

59. Because of data scarcity, we use a five-year rolling window around the year where we estimate the technology. In Section VI and in the Appendix, we discuss the estimation routine in further detail and show the robustness of our findings with the baseline technology to different technological specifications.

omitted price variable bias). Both issues are of course not independent, and we rely on methods that aim to deliver consistently estimated output elasticities, dealing with them adequately.

We follow the literature and control for the simultaneity and selection bias, inherently present in the estimation of equation (23), and rely on a control function approach, paired with a law of motion for productivity, to estimate the output elasticity of the variable input.[60] This method accounts for the fact that the variable factor of production V adjusts in response to a productivity shock, whereas the fixed factor K does not react to contemporaneous shocks to productivity, but it is correlated with the persistent productivity term. This requires us to restrict the production function to a particular class to guarantee that the coefficients of interest—which determines the output elasticity—are identified.

1. Control Function. We build on the insight from Olley and Pakes (1996) that (unobserved) productivity ω_{it} can be expressed as an (unknown) function of the firm's state variables and observables. This is obtained by considering input (or investment) demand, and inverting out for productivity to yield:

$$\omega_{it} = h_t(d_{it}, k_{it}, z_{it}), \tag{25}$$

where d_{it} is the control variable. We consider two cases: a variable input in production (in our case COGS, v), and investment (i), and Ackerberg et al. (2007) provide an excellent treatment of the two types of control variables. z_{it} captures output and input market factors that generate variation in factor demand (for input d) across firms, conditional on the level of productivity and capital. The latter is critical to allow for imperfectly competitive product markets when estimating production functions. Standard approaches in the literature on production function estimation are restricted to either perfect competition or models of common markups (monopolistic competition paired with CES demand).

60. We estimate the production function, by industry, over an unbalanced panel to deal with the nonrandom exit of firms, which Olley and Pakes (1996) found to be important. However, the source of the attrition in the Compustat data is likely to be different than in traditional plant-level manufacturing data sets—that is, firms drop out of the data due to both exit and mergers and acquisitions, so the sign of the bias induced by the selection is ambiguous. We are, however, primarily interested in estimates of the variable output elasticity, while the selection bias is expected to impact the capital coefficient more directly.

Instead, we rely on De Loecker and Warzynski (2012) and De Loecker et al. (2016) to allow for imperfect competition in product markets and thus markup heterogeneity across firms. In practice this amounts to allowing for input demand shifters that move around the optimal amount of a variable input, conditional on a firm's productivity and capital stock.

Regardless of which control variable is used, this method relies on a so-called two-stage approach. In the first stage, the measurement error and unanticipated shocks to output are purged using a nonparametric projection of output on the inputs and the control variable.

In the case of a static control, $d_{it} = v_{it}$, this is given by:

$$y_{it} = \phi_t(v_{it}, k_{it}, z_{it}) + \epsilon_{it}. \tag{26}$$

The output elasticity is obtained by constructing moments of the productivity shock, which is obtained by considering a productivity process given by $\omega_{it} = g(\omega_{it-1}) + \xi_{it}$. It gives rise to the following moment condition to obtain the industry-year-specific output elasticity:

$$\mathbb{E}\left(\xi_{it}(\theta_t)\begin{bmatrix} v_{it-1} \\ k_{it} \end{bmatrix}\right) = 0, \tag{27}$$

where $\xi_{it}(\theta_t)$ is obtained by projecting productivity $\omega_{it}(\theta_t)$ on its lag $\omega_{it-1}(\theta_t)$, with $\theta_t = \{\theta_t^V, \theta_t^K\}$, where productivity is in turn obtained from $\phi_{it} - \theta_t^V v_{it} - \theta_t^K k_{it}$, using the estimate ϕ_{it} from the first-stage regression. This approach identifies the output elasticity of a variable input under the assumption that the variable input use responds to productivity shocks but the lagged values do not, and that lagged variable input use is correlated with current variable input use, through serially correlated input and output market conditions, captured in z_{it}. In the case of PF2, an additional moment identifies the output elasticity of the overhead (SG&A) input, $\mathbb{E}(\xi_{it}(\theta_t)x_{it}) = 0$.

In the case of the Olley-Pakes approach, we can identify and estimate the output elasticity using a simple nonlinear regression:

$$y_{it} = \theta_t^V v_{it} + \phi_t(i_{it}, k_{it}, z_{it}) + \epsilon_{it}, \tag{28}$$

and rely on the identification arguments made in Ackerberg, Caves, and Frazer (2015)—that is, if the variable input bundle v (COGS) is nondynamic and chosen at $t - b$ $(0 < b < 1)$, while

the investment decision is made at t, allowing for productivity shocks to hit the firm between these two subperiods. This approach has the advantages that it is simple to implement and does not require us to consider the subsequent second stage. Compared with the static control, discussed above, the investment policy function needs to be increasing in productivity (conditional on capital and variables captured by z), and the specification adopted here limits the scope of strategic interaction among firms.

We consider both controls (COGS and investment) and find very similar results for the estimated output elasticities. Below we plot the two series, using the static and dynamic control variable, and we aggregate the industry-year-specific output elasticities of COGS using industry sales.

A.B. Units

As pointed out in De Loecker and Goldberg (2014) standard production data, whether it is Compustat or census data, records revenue and expenditures, rather than physical production and input use (with the exception of a few manufacturing industries). In the presence of product differentiation (be it through physical attributes or location) an additional source of endogeneity presents itself through unobserved output and input prices. This has been the topic of recent research, for a recent treatment see De Loecker et al. (2016). A first observation is that the error term, ϵ_{it}, will generally contain output and input prices (scaled by the relevant technology parameters). De Loecker et al. (2016) show that the correlation of input expenditures with this error yields biased estimates of the output elasticity. However, in their setting physical output quantities are observed, and the unobserved input prices, reflecting differentiation, are the only source of the price error. We do not observe output price variation, and we are therefore left with the following structural error term:

$$\omega_{it} + p_{it} - \theta_t^V p_{it}^v - \theta_t^K p_t^K, \tag{29}$$

where we let the user cost of capital be industry-time specific, but input prices potentially vary across firms reflecting variation in quality, location, and other exogenous factors. We follow De Loecker et al. (2016) and let the wedge between the output and input price (scaled by the output elasticity) be a function of the

demand shifters and productivity difference.[61] In the case of Olley and Pakes, the inclusion of the variable z should therefore capture the relevant output and input market forces that generate differences in output and input price. Of course, productivity differences that influence the wedge between output and input prices are automatically captured by the inclusion of the control function. Note that not observing output prices has the perhaps unexpected benefit that output price variation absorbs input price variation, thus eliminating part of the variation in the error term. In the extreme case, we are left with just the productivity unobservable, and this puts us back in the standard framework introduced above.

Under the alternative DGP, where the static control (COGS) is used and where the output elasticity is identified in the second stage, we follow De Loecker et al. (2016). The main difference lies in the fact that we cannot rely on observed output prices, and we therefore have to rely on constructed measures of market share (at various levels of aggregation) to eliminate the variation in the price error wedge.

In practice, we consider market share, measured at various levels of aggregation (two, three, and four digit), to take into account additional variation in output and input markets. As discussed in De Loecker et al. (2016) this is an exact control when output prices, conditional on productivity, reflect input price variation, and when demand is of the (nested) logit form. We have subjected our analysis to a host of different specifications of the production function (such as the translog production function), and we find similar results for the estimated output elasticities. As discussed in the main text, the main findings on aggregate markups are furthermore not sensitive to the use of a common time-invariant calibrated output elasticity of 0.85. We also considered an alternative specification, including SG&A as a factor of production, and document a comparable rise in aggregate markups.

Finally, there are a host of possible measurement error and endogeneity concerns with any single specification we could consider. We do not attempt to provide the one final set of output elasticities for all sectors of the U.S. economy, using the Compustat data. Rather we consider a variety of specifications, and

61. See De Loecker et al. (2016) for a microfoundation for this, and application. An alternative is De Loecker (2011) and specifying a specific demand system. This, however, limits the scope of markup heterogeneity, across firms and time, and we are precisely interested in describing markups in a flexible fashion.

show that the main facts we are interested in (under the maintained year-sector specific Cobb-Douglas production function) are not sensitive to these. The aggregate markup can be expressed in terms of the potential bias $\psi_{st} = \hat{\theta}_{st} - \theta_{st}$ in the production function coefficient:

$$\begin{aligned}\mu_t &= \sum_i m_{it}\hat{\mu}_{it} - \sum_i m_{it}\psi_{st} \\ &= \hat{\mu}_t - \sum_s m_{st}\psi_{st}.\end{aligned}$$

We have no prior belief that there is a particular correlation between the weight of an industry in the economy, or in the sample, and the bias introduced by either the simultaneity, selection, or omitted price variable bias.

Appendix B: Data: Summary Statistics

B.A. Compustat

We obtain firm-level financial variables of all U.S.-incorporated publicly listed companies active at any point during the period 1950–2016. We access the Compustat North America Fundamentals Annual (through WRDS) and download the annual accounts for all companies. The results in this article are obtained with a download on March 25, 2018. We keep unique records for each firm and assign a firm to a unique two-digit industry, as reported. We exclude firms that do not report an industry code. All financial variables are deflated with the appropriate deflators. The main results, unless reported otherwise, rely on the sample of firms over the period 1950–2016, where we eliminate firms with reported cost-of-goods to sales and SG&A to sales ratios in the top and bottom 1%, where the percentiles are computed for each year separately. Our results are invariant to trimming up to 5% (bottom and top).[62] As such, a firm-year observation requires information on both sales and COGS, two essential ingredients to measure markups. Appendix Table B.1 below presents a few basic summary statistics for a few leading variables used in our analysis (sales, COGS, capital, wage bill, employment, and SG&A), for

62. Rather than winsorizing the tails of the distribution, we find similar results when doing structural error correction to purge the measurement error from sales using the specification of the control function in equation (26).

TABLE B.1
SUMMARY STATISTICS (1955–2016)

		Sample A		
	Acronym, var.	Mean	Median	No. obs
Sales	*SALE*, *PQ*	1,922,074	147,806	247,644
Cost of goods sold	*COGS*, *V*	1,016,550	55,384	247,644
Capital stock	*PPEGT*, *K*	1,454,210	57,532	247,644
SG&A	*XSG&A*, *X*	342,805	29,682	247,644
Wage bill	*XLR*, *WL*	1,093,406	130,486	28,116
Employment	*EMP*, *L*	8,363	863	221,121
		Sample B		
	Acronym, var.	Mean	Median	No. obs
Sales	*SALE*, *PQ*	5,894,779	578,912	28,116
Cost of goods sold	*COGS*, *V*	2,970,693	195,087	28,116
Capital stock	*PPEGT*, *K*	5,193,319	345,592	28,116
SG&A	*XSG&A*, *X*	926,542	78,487	28,116
Wage bill	*XLR*, *WL*	1,093,406	130,486	28,116
Employment	*EMP*, *L*	24,861	4,522	25,527

Notes: Thousands US$ deflated using the GDP deflator with base year 2010. For each variable we list the Compustat acronym, the associated notation (in levels) used throughout the manuscript.

two samples. Sample A, observations with information on sales, COGS, and SG&A; and Sample B, observations with information on the wage bill.

B.B. Economic Censuses

The focus of our analysis of the census data is on manufacturing (NAICS codes 31-32-33), retail (NAICS codes 44-45), and wholesale (NAICS code 42). In 2012, manufacturing consists of about 297,000 establishments, retail of about 1,060,000 establishments, and wholesale of about 420,000 establishments. These establishments aggregate into about 650,000 retail firms, about 314,000 wholesale firms, and about 250,000 manufacturing firms. Together these three sectors make up a little over 20% of U.S. GDP. In principle, each economic census spans the universe of every single employer establishment in its sector, across the size distribution; only nonemployer establishments (sole proprietorships with no employees) are omitted.

The other censuses that we do not use are the Census of Services, the Census of Construction Services; the Census of Mining; the Census of Transportation, Communications, and Utilities; the

Census of Finance, Insurance, and Real Estate; and the Census of Auxiliary Establishments.[63]

The data are organized around the most discrete unit of production in the microdata, an "establishment," which is a single physical plant. Establishments can be aggregated to the EIN level (Employer Identification Number, the most discrete legal unit of production; an EIN is a unique tax ID associated with a distinct legal entity), and higher up to the firm level (major corporations are usually collections of EINs, which in turn are collections of multiple establishments). The microdata associate each establishment with an EIN and a firm ID: the EIN is considered part of the firm if the firm has complete or majority ownership of the EIN.

Perhaps the most common way of defining "firm" in the recent firm heterogeneity literature is to say that all of a firm's establishments in a given four-digit SIC industry (roughly equivalent to a six-digit NAICS industry) are a distinct firm (this is the approach taken by Hsieh and Klenow 2009; Autor et al. 2020; and others). Under this definition, Walmart's establishments listed as, for example, SIC 5411 (Retail - Grocery Stores) are one firm, and Walmart's establishments listed in SIC 5412 (Retail - Convenience Stores) are a separate firm. Our preferred default approach is to define "firm" as all of the firm's establishments in a single sector census (e.g., all of Walmart's firms in all of retail, NAICS codes 44-45, are a single firm).

Notice that even though we do not have information on overhead directly, one can obtain multiple sources of information about overhead costs in the census data: (i) Census flags the auxiliary establishments of multiunit firms and links them to other establishments of the same firm. Auxiliary establishments include headquarters, other facilities mainly engaged in general management functions, and facilities that mainly engage in R&D. Census has taken a systematic approach to identifying and flagging auxiliary establishments across most sectors of the economy since 1997. (ii) Census conducts various business surveys that elicit information about various types of overhead. For example, the 2012 Annual Survey of Manufactures includes questions about software expenses, the cost of purchased communication services, advertising and promotional expenses, and the cost of purchased professional and technical services. As a second example, the Survey of Industrial R&D collects data on research and development

63. Note that the Census of Auxiliary Establishments includes many corporate support units that do not directly face customers or take in revenue.

expenses for all large firms and a sample of smaller ones. Although the data are scattered across a variety of sources and databases, there is great potential in constructing firm-level measures of overhead costs using the census data, and combining it with the markup analysis. This is left for future work and lies beyond the scope of this paper.[64]

APPENDIX C: THE DISTRIBUTION OF MARKUPS WITH TECHNOLOGY PF2

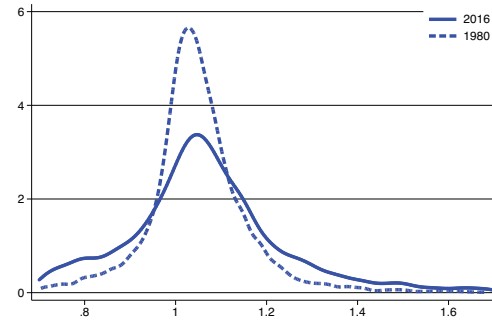


FIGURE C.1

Distribution of Markups μ_{it}: Kernel Density Plots (Unweighted)

64. We thank an anonymous referee for pointing this out to us and for providing a careful and detailed discussion on the various data sources at Census to measure overhead costs.

APPENDIX D: AN ALTERNATIVE PRODUCTION FUNCTION: UNBUNDLING COGS

We have used a bundle of inputs (COGS) as well as the wage bill only, which is included in COGS, to estimate markups. We now unbundle COGS into the wage bill and materials, which we calculate as the residual of COGS minus the wage bill, denoted by $\tilde{M}$. In doing so, we face the well-known limitation in the Compustat sample that the wage bill is reported only by a small number of firms.

To compute the markup using the first-order condition on labor, we require an output elasticity of labor. This elasticity is obtained as before by estimating the production function. However, now we have to distinguish between labor and intermediate inputs, taken together in the variable input bundle V, and this requires a modeling choice as to how intermediate inputs enter the production function. We consider a fixed-proportion (Leontief) technology in the intermediate variable. This is the case considered in De Loecker and Scott (2016), Ackerberg, Caves, and Frazer (2015), and Gandhi, Navarro, and Rivers (2011) and avoids the potential identification issues surrounding intermediate inputs in the classical setting. Consider:

$$Q_{it} = \min\{\theta^M m_{it}, L_{it}^{\theta_L} K_{it}^K \Omega_{it}\}. \tag{30}$$

We estimate this production function by sector, for each year with sector fixed effects. There are not enough observations to reliably estimate the production function by sector/year. The main insight is that we do not need to observe intermediate inputs to estimate the production function, but instead we project gross output on labor and capital. To compute the markup, however, we have to include intermediates because the marginal cost of production requires the appropriate increase in intermediate inputs when increasing labor. We derive the markup from the first-order condition accounting for the fact that the nondifferentiable technology equation (30) requires input choices in fixed proportions:

$$\mu_{it}(L) = \frac{1}{\mu_{it}^{-1} + \frac{P_t^M m_{it}}{P_{it} Q_{it}}}, \tag{31}$$

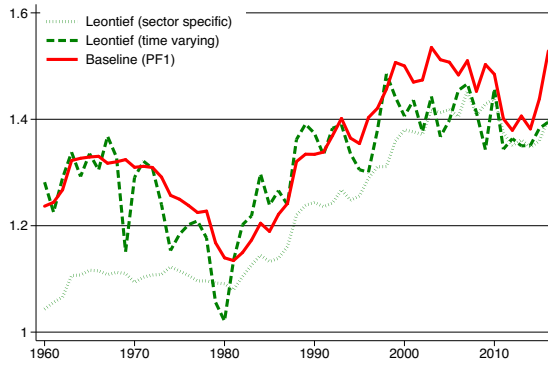


FIGURE D.1

Industry-Specific Average Markups Using Labor Cost

where μ_{it} is obtained with the standard formula and the output elasticity is the estimated labor coefficient, that is, $\mu_{it} = \theta^L \frac{S_{it}}{wL_{it}}$. We compute the material expenditure by netting the wage bill from COGS.

We report in Appendix Figure D.1 the average markup: green dashed lines (in print; green solid and long-dashed lines online): Leontief technology (sector specific and time varying respectively); solid red line (in print; dotted red in color version online): baseline markup (PF1) for the selected sample with data on the wage bill.

APPENDIX E: MARKUP DISTRIBUTION: AUTOREGRESSIVE PROCESS

To capture some properties of the process that governs the evolution of markups, we assume the following autoregressive process for our measure of markups (with the conventional production function) as well as for the data on sales and employment:

$$z_{it} = \rho x_{it-1} + \varepsilon_{it}, \quad z \in \{\log \mu, \log S, \log L\}. \tag{32}$$

Appendix Figure E.1 shows the evolution of the cross-sectional standard deviation of the shocks in the markup, sales,

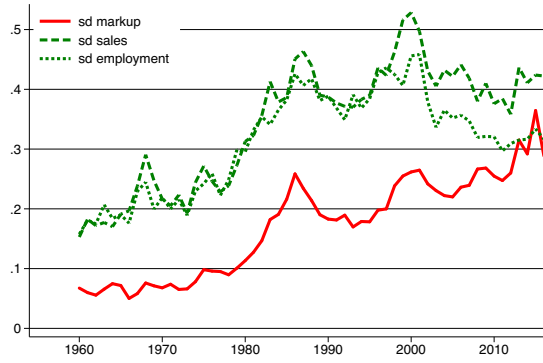


FIGURE E.1

The Evolution of the Standard Deviation of Markups, Sales, and Employment (1960–2016).

AR(1) in logs on their lag with year and industry fixed effects. The estimated persistence is 0.84.

and employment processes. Starting in 1980, there is clearly a sharp rise in the standard deviation of the markup μ and a more moderate increase in that of sales S. Interestingly, there is much less of an increase in the standard deviation of employment L. If anything, there is a decline from 2000 onward. This is because the increase in the standard deviation of markups is precisely driven by the increase in the wedge between the volatility of sales (increasing) and inputs, in this case labor (fairly constant and then decreasing).[65]

This increasing wedge is consistent with the evidence in Decker et al. (2014) that the shock process itself has not changed much but that the transmission of the shocks to inputs (labor) has.

65. As robustness checks, we also included higher-order terms of the persistence and find that those are not important.

APPENDIX F: AGGREGATE PROFITS

We compare aggregate profits with the markup measure τ proposed in Traina (2018). See Appendix Figure F.1.

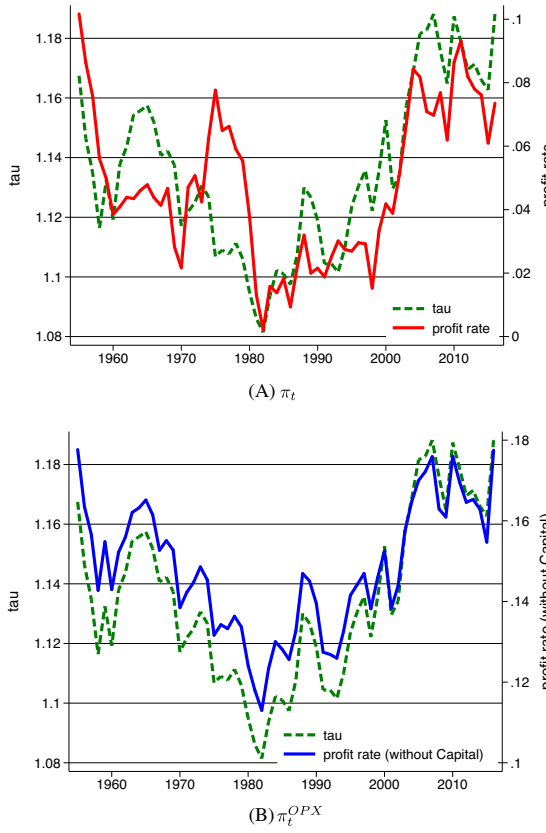


FIGURE F.1

τ versus Profit Rates

KATHOLIEKE UNIVERSITEIT LEUVEN, NATIONAL BUREAU OF ECONOMIC RESEARCH, AND CENTRE FOR ECONOMIC POLICY RESEARCH
UNIVERSITAT POMPEU FABRA BARCELONA (CATALAN INSTITUTION FOR RESEARCH AND ADVANCED STUDIES, GRADUATE SCHOOL OF ECONOMICS)
HARVARD UNIVERSITY

SUPPLEMENTARY MATERIAL

An Online Appendix for this article can be found at *The Quarterly Journal of Economics* online. Data and code replicating tables and figures in this article can be found in De Loecker, Eeckhout, and Unger (2020), in the Harvard Dataverse, doi: 10.7910/DVN/5GH8XO.

REFERENCES

Ackerberg, Daniel, C. Lanier Benkard, Steven Berry, and Ariel Pakes, "Econometric Tools for Analyzing Market Outcomes," *Handbook of Econometrics*, 6 (2007), 4171–4276.
Ackerberg, Daniel A., Kevin Caves, and Garth Frazer, "Identification Properties of Recent Production Function Estimators," *Econometrica*, 83 (2015), 2411–2451.
Antràs, Pol, "Is the U.S. Aggregate Production Function Cobb-Douglas? New Estimates of the Elasticity of Substitution," *Contributions to Macroeconomics*, 4 (2004).
Atkeson, Andrew, and Ariel Burstein, "Pricing-to-Market, Trade Costs, and International Relative Prices," *American Economic Review*, 98 (2008), 1998–2031.
Autor, David, David Dorn, Lawrence F. Katz, Christina Patterson, and John Van Reenen, "The Fall of the Labor Share and the Rise of Superstar Firms," *Quarterly Journal of Economics*, 135 (2020), 645–709.
Baqaee, David Rezza, and Emmanuel Farhi, "The Macroeconomic Impact of Microeconomic Shocks: Beyond Hulten's Theorem," *Econometrica*, 87 (2019), 1155–1203.
———, "Productivity and Misallocation in General Equilibrium," *Quarterly Journal of Economics*, 135 (2020), 105–163.
Barkai, Simcha, "Declining Labor and Capital Shares," Mimeo, Chicago Booth, 2017.
Basu, Susanto, "Are Price-Cost Markups Rising in the United States? A Discussion of the Evidence," *Journal of Economic Perspectives*, 33 (2019), 3–22.
Basu, Susanto, and John G. Fernald, "Returns to Scale in US Production: Estimates and Implications," *Journal of Political Economy*, 105 (1997), 249–283.
Berger, David, and Joseph Vavra, "Shocks vs. Responsiveness: What Drives Time-Varying Dispersion?," NBER Technical report, 2017.
Berry, Steven, James Levinsohn, and Ariel Pakes, "Automobile Prices in Market Equilibrium," *Econometrica*, 63 (1995), 841–890.
Brennan, Jordan, "Rising Corporate Concentration, Declining Trade Union Power, and the Growing Income Gap: American Prosperity in Historical Perspective," Mimeo, Levys Economics Institute, 2016.
Bresnahan, Timothy F., "Empirical Studies of Industries with Market Power," *Handbook of Industrial Organization*, 2 (1989), 1011–1057.

Brown, Gregory, and Nishad Kapadia, "Firm-Specific Risk and Equity Market Development," *Journal of Financial Economics*, 84 (2007), 358–388.
Burnside, Craig, "Production Function Regressions, Returns to Scale, and Externalities," *Journal of Monetary Economics*, 37 (1996), 177–201.
Campa, Jose Manuel, and Linda S. Goldberg, "Exchange Rate Pass-Through into Import Prices," *Review of Economics and Statistics*, 87 (2005), 679–690.
Cooper, Russell, and João Ejarque, "Financial Frictions and Investment: Requiem in Q," *Review of Economic Dynamics*, 6 (2003), 710–728.
Davis, Steven J., and John Haltiwanger, "Labor Market Fluidity and Economic Performance," NBER Technical report, 2014.
Davis, Steven J., John Haltiwanger, Ron Jarmin, and Javier Miranda,"Volatility and Dispersion in Business Growth Rates: Publicly Traded Versus Privately Held Firms," In *NBER Macroeconomics Annual 2006*, (Cambridge, MA: MIT Press, 2007), Vol. 21, 107–180.
De Loecker, Jan, "Product Differentiation, Multiproduct Firms, and Estimating the Impact of Trade Liberalization on Productivity," *Econometrica*, 79 (2011), 1407–1451.
De Loecker, Jan, and Jan Eeckhout, "Global Market Power," NBER Working Paper, 2018a.
———, "Some Thoughts on the Debate about (Aggregate) Markup Measurement," Mimeo, Universitat Pompeu Fabra Working Paper, 2018b.
De Loecker, Jan, Jan Eeckhout, and Simon Mongey, "Quantifying Market Power," Mimeo, 2018.
De Loecker, Jan, Jan Eeckhout, and Gabriel Unger, "Replication Data for: The Rise of Market Power and the Macroeconomic Implications," (2020), Harvard Dataverse: doi: 10.7910/DVN/5GH8XO.
De Loecker, Jan, and Pinelopi Koujianou Goldberg, "Firm Performance in a Global Market," *Annual Review of Economics*, 6 (2014), 201–227.
De Loecker, Jan, Pinelopi K. Goldberg, Amit K. Khandelwal, and Nina Pavcnik, "Prices, Markups and Trade Reform," *Econometrica*, 84 (2016), 445–510.
De Loecker, Jan, and Paul T. Scott, "Estimating Market Power Evidence from the US Brewing Industry" NBER Technical report, 2016.
De Loecker, Jan, and Frederic Michel Patrick Warzynski, "Markups and Firm-Level Export Status," *American Economic Review*, 102 (2012), 2437–2471.
Decker, Ryan, John Haltiwanger, Ron Jarmin, and Javier Miranda, "The Secular Decline in Business Dynamism in the US," Mimeo, University of Maryland, 2014.
Doidge, Craig, G. Andrew Karolyi, and René M Stulz, "The US Listing Gap," *Journal of Financial Economics*, 123 (2017), 464–487.
Edmond, Chris, Virgiliu Midrigan, and Daniel Yi Xu, "Competition, Markups, and the Gains from International Trade," *American Economic Review*, 105 (2015), 3183–3221.
———, "How Costly Are Markups?," NBER Technical report, 2019.
Eeckhout, Jan, and Xi Weng, "The Technological Origins of the Decline in Labor Market Dynamism," Mimeo, 2017.
Elsby, Michael W. L., Bart Hobijn, and Ayşegül Şahin, "The Decline of the US Labor Share," *Brookings Papers on Economic Activity*, 2013 (2013), 1–63.
Engbom, Niklas, "Firm and Worker Dynamics in an Aging Labor Market," Technical report, 2017.
Fallick, Bruce, Charles Fleischman, and Jonathan Pingle, "The Effect of Population Aging on the Aggregate Labor Market," in *Labor in the New Economy*, Katharine G. Abraham, James R. Spletzer, and Michael Harper, eds. (Chicago: University of Chicago Press, 2010), 377–417.
Fama, Eugene F., and Kenneth R. French, "New Lists: Fundamentals and Survival Rates," *Journal of Financial Economics*, 73 (2004), 229–269.
Foster, Lucia, John Haltiwanger, and Chad Syverson, "Reallocation, Firm Turnover, and Efficiency: Selection on Productivity or Profitability?," *American Economic Review*, 98 (2008), 394–425.

Gabaix, Xavier, "The Granular Origins of Aggregate Fluctuations," *Econometrica*, 79 (2011), 733–772.
Gabaix, Xavier, and Augustin Landier, "Why Has CEO Pay Increased So Much?," *Quarterly Journal of Economics*, 123 (2008), 49–100.
Ganapati, Sharat, Joseph S. Shapiro, and Reed Walker, "The Incidence of Carbon Taxes in U.S. Manufacturing: Lessons from Energy Cost Pass-Through," Yale University Technical report 2018.
Gandhi, Amit, Salvador Navarro, and David A. Rivers, "On the Identification of Production Functions: How Heterogeneous is Productivity?," Mimeo, 2011.
Gao, Xiaohui, Jay R. Ritter, and Zhongyan Zhu, "Where Have All the IPOs Gone?," *Journal of Financial and Quantitative Analysis*, 48 (2013), 1663–1692.
Gollin, Douglas, "Getting Income Shares Right," *Journal of Political Economy*, 110 (2002), 458–474.
Grassi, Basile, "IO in IO: Competition and Volatility in Input-Output Networks," Unpublished Manuscript, Bocconi University, 2017.
Grullon, Gustavo, Yelena Larkin, and Roni Michaely, "Are US Industries Becoming More Concentrated?," Unpublished Working Paper, 2016.
Gutiérrez, Germán, and Thomas Philippon, "Declining Competition and Investment in the US," NBER Technical report, 2017.
———, "How EU Markets Became More Competitive than US Markets: A Study of Institutional Drift," NBER Working Paper no. 24700, 2018.
Hall, R. E., "The Relation between Price and Marginal Cost in U.S. Industry," *Journal of Political Economy*, 96 (1988), 921–947.
Hall, Robert E., "New Evidence on the Markup of Prices over Marginal Costs and the Role of Mega-Firms in the US Economy," NBER Technical report, 2018.
Haltiwanger, John C., "Measuring and Analyzing Aggregate Fluctuations: The Importance of Building from Microeconomic Evidence," *Federal Reserve Bank St. Louis Review*, 79 (1997), 55–77.
Hartman-Glaser, Barney, Hanno Lustig, and Mindy X. Zhang, "Capital Share Dynamics When Firms Insure Workers," NBER Technical report, 2016.
Hsieh, Chang-Tai, and Peter J. Klenow, "Misallocation and Manufacturing TFP in China and India," *Quarterly Journal of Economics*, 124 (2009), 1403–1448.
Hyatt, Henry R., and James R. Spletzer, "The Recent Decline in Employment Dynamics," *IZA Journal of Labor Economics*, 2 (2013), 5.
Kaplan, Greg, and Sam Schulhofer-Wohl, "Understanding the Long-Run Decline in Interstate Migration," NBER Technical report, 2012.
Karabarbounis, Loukas, and Brent Neiman, "The Global Decline of the Labor Share," *Quarterly Journal of Economics*, 129 (2013), 61–103.
———, "Accounting for Factorless Income," NBER Technical report, 2018.
Karahan, Fatih, Benjamin Pugsley, and Ayşegül Şahin, "Demographic Origins of the Startup Deficit," Mimeo, NY Fed, 2016.
Kehrig, Matthias, "The Cyclicality of Productivity Dispersion," U.S. Census Bureau Center for Economic Studies Paper No. CES-WP-11-15, 2011.
Kehrig, Matthias, and Nicolas Vincent, "Growing Productivity without Growing Wages: The Micro-Level Anatomy of the Aggregate Labor Share Decline," Mimeo, Duke University, 2017.
Keller, Wolfgang, and Stephen Yeaple, "Multinational Enterprises, International Trade, and Productivity Growth: Firm-Level Evidence from the United States," *Review of Economics and Statistics*, 91 (2009), 821–831.
Koh, Dongya, Raul Santaeulalia-Llopis, and Yu Zheng, "Labor Share Decline and Intellectual Property Products Capital," Mimeo, Washington University, 2017.
Koujianou Goldberg, Pinelopi, and Rebecca Hellerstein, "A Structural Approach to Identifying the Sources of Local Currency Price Stability," *Review of Economic Studies*, 80 (2012), 175–210.
Mertens, Matthias, "Micro-Mechanisms behind Declining Labor Shares: Market Power Production Processes, and Global Competition," Mimeo, Halle Institute for Economic Research, 2019.

Morlacco, Monica, "Market Power in Input Markets: Theory and Evidence from French Manufacturing," Yale University Technical report, 2017.
Nevo, Aviv, "Measuring Market Power in the Ready-to-Eat Cereal Industry," *Econometrica*, 69 (2001), 307–342.
Nishida, Mitsukuni, Amil Petrin, Martin Rotemberg, and T. White, "Are We Undercounting Reallocation's Contribution to Growth?," U.S. Census Technical report, 2017.
Olley, G. Steven, and Ariel Pakes, "The Dynamics of Productivity in the Telecommunications Equipment Industry," *Econometrica*, 64 (1996), 1263–1297.
Rubens, Michael, "Monopsony Power and Factor-Biased Technology Adoption," Mimeo, University of Leuven, 2019.
Smith, Matthew, Danny Yagan, Owen Zidar, and Eric Zwick, "Capitalists in the Twenty-first Century," UC Berkeley and University of Chicago Working Paper, 2017.
Syverson, Chad, "Market Structure and Productivity: A Concrete Example," *Journal of Political Economy*, 112 (2004), 1181–1222.
———, "Macroeconomics and Market Power: Context, Implications, and Open Questions," *Journal of Economic Perspectives*, 33 (2019), 23–43.
Traina, James, "Is Aggregate Market Power Increasing? Production Trends using Financial Statements," Mimeo, Chicago Booth, 2018.