Monopsony in the U.S. Labor Market[*]

Brad Hershbein
W.E. Upjohn Institute

Claudia Macaluso
FRB Richmond

Chen Yeh
FRB Richmond

December 31, 2019

Abstract

This paper quantifies whether the U.S. labor market is characterized by employer market power and whether the degree of employer market power has increased over time. We find that the vast majority of U.S. manufacturing plants operate in a monopsonistic environment and, at least since the early 2000s, the labor market in U.S. manufacturing has become more monopsonistic. To reach this conclusion, we exploit rich administrative data for U.S. manufacturers and estimate plant-level markdowns — the ratio between a plant's marginal revenue product of labor and its wage. In a competitive labor market, markdowns would be equal to unity. Instead, we find substantial deviations from perfect competition, as markdowns average at 1.53. This result implies that a worker employed at the average manufacturing plant earns 65 cents on each dollar generated on the margin. Furthermore, we document a substantial amount of dispersion in markdowns across plants, even within detailed industries. To investigate long-term trends in employer market power, we propose a novel measure for the aggregate markdown that is consistent with aggregate wedges and also incorporates the local nature of labor markets. We find that the aggregate markdown decreased from the late 1970s up to the early 2000s, but has been sharply increasing afterward. When we compare often-used indexes of employment concentration with our markdown estimates, we find a dissimilar evolution over time. This weak relationship cautions against interpreting employment concentration as a proxy for employer market power.

Key words: Monopsony, labor market power, markdowns, secular trends

JEL: E2, J2, J3, J42

[*] Contact: Hershbein (hershbein@upjohn.org), Macaluso (claudia.macaluso@rich.frb.org) and Yeh (chen.yeh@rich.frb.org). We thank Adrien Auclert, Christian Bayer, Steven J. Davis, Jan Eeckhout, Grey Gordon, Lisa Kahn, Loukas Karabarbounis, Ayşegül Şahin, Pierre-Daniel Sarte, Felipe Schwartzman, Chad Syverson and Nicolas Trachter in addition to the audiences at many seminars and conferences, including SITE "Micro and Macro of Labor markets", NBER 2019 SI CRIW and MP groups, Barcelona Summer Forum EGF 2019, FRB Richmond "Market Structure and the Macroeconomy", University of Oslo-CEPR "Micro-consistent Macro", and many others. This research was conducted while Claudia Macaluso and Chen Yeh were Special Sworn Status researchers at the U.S. Census Bureau. We also thank the hospitality of the University of Minnesota, and the Chicago, Minnesota, and Federal Reserve Board Federal Research Data Centers, where part of this research was conducted. Any opinions and conclusions expressed herein are those of the authors and do not necessarily represent the views of the U.S. Census Bureau, the Federal Reserve Bank of Richmond, or the Federal Reserve System. All results have been reviewed to ensure that no confidential information is disclosed. First version: May 15, 2018.

1 Introduction

Is the U.S. labor market perfectly competitive? In perfectly competitive labor markets, marginal revenue products of labor are equal to workers' wages meaning that every dollar generated on the margin is paid to workers. Despite this assumption's modeling convenience, does this benchmark accurately describe the U.S. labor market? Wedges between marginal revenue products of labor and wages may constitute evidence of monopsony and suggest a departure from allocative efficiency. In this paper, we provide estimates of these wedges — "markdowns" — across U.S. manufacturing plants from 1976 to 2014. Specifically, we show that (i) the U.S. manufacturing labor market is characterized by significant markdowns, consistent with employer market power, and (ii) the degree of this market power decreased between the late 1970s and the early 2000s, but increased sharply afterwards.

Quantifying employers' market power and understanding its dynamics across employers and over time is fundamental to devise appropriate policy responses. Reliable evidence on employer market power is particularly relevant to evaluate policies that directly affect workers' compensation and mobility, such as changes in the minimum wage. Similarly, when assessing regulatory limits to the growth of large firms, it is helpful to consider to what extent such firms are able to compensate labor below their marginal revenue products. Far from being only theoretical possibilities, policy makers have recently actively considered these policies to mitigate a perceived increase in employers' market power (cfr. FTC, 2018).[1] While this rise in employer market power can be plausibly connected to several labor market trends, direct measures of employer market power are not available to inform the current policy debate. Our paper responds precisely to this gap. We estimate plant-level markdowns for the whole U.S. manufacturing sector and study their relationship with employer size, age, scope, and productivity, and the evolution of aggregate markdowns over time.

Our analysis of monopsony in the labor market starts from estimating and characterizing the distribution of plant-level markdowns. In our baseline framework, firms internalize a finitely elastic labor supply curve, thus operating in a monopsonistic environment. Without imposing further restrictions on the labor supply curve, we interpret gaps between the output elasticity of labor and its revenue share as market power in output and/or labor markets (markups and/or markdowns). Under the assumption that at least one observable input is flexible, markups and markdowns can be identified and estimated separately. To implement this insight empirically, we closely follow the production function estimation approach from the industrial organization (IO) literature and use comprehensive administrative data for the U.S. manufacturing sector.

Our results indicate that labor markets in U.S. manufacturing are far from perfectly competitive. Estimated markdowns are considerably larger than unity and diverse across plants, even within detailed industries. The average plant's marginal revenue product of labor is 53 percent higher than its wage, implying that a worker employed in the average plant receives about 65 cents on a dollar generated on the margin. Furthermore, we document a substantial amount of dispersion across plants even within 3-digits NAICS industries, with

[1] See the Federal Trade Commission Hearing #3: Multi-Sided Platforms, Labor Markets and Potential Competition on October 15–17, 2018.

an average within-industry interquartile range of 61.6 percent. Investigating the sources of heterogeneity in markdowns, we find that markdowns are positively correlated with size, and a firm's sectoral and geographical scope. Hence, our results lend support to the view that size dependency is quantitatively relevant to assess the welfare implications of labor market power. We conclude that the average U.S. manufacturing plant operates in a monopsonistic environment and the variation in markdowns across plants is mainly linked to idiosyncratic factors rather than industry-wide ones such as legacy structure or sector-specific regulations.

With estimates of micro-level markdowns in hand, we focus on documenting the dynamics of macro-level markdowns since 1977. There is no uncontested framework that delivers a clear aggregation rule for markdowns. To deal with the issue of aggregation, we propose a measure of an "aggregate markdown" that satisfies two requirements. First, aggregate markdowns and markups reflect *aggregate* wedges which are gaps that a fictional representative firm would face. This interpretation has the advantage that no specific market structure for labor or output needs to be imposed in order to shed light on aggregation. Second, aggregate markdowns need to take the local nature of labor markets into account, which is consistent with evidence on the cost of distance during job search. In the end, we show that aggregation occurs through sales-weighted harmonic averages, where weights are adjusted for heterogeneity in output elasticities. Our measure of the aggregate markdown displays a U-shaped evolution over time, decreasing between 1977 and 2002, and sharply increasing afterwards. All in all, the data supports the hypothesis of an increase in the degree of monopsony in the U.S. manufacturing labor market since the early 2000s.

We identify markdowns through the so-called "production approach". This approach has several advantages. First of all, we do not need to take a stand on the sources of employer market power. In fact, we explicitly show that our approach is consistent with a broad range of monopsony models. On the other hand, we only need to impose a functional form on a firm's production function. To be as flexible as possible, we assume that production functions are translog —- a second-order approximation to *any* arbitrary production function. A second benefit of the production approach is that it remains valid regardless of the assumptions made on other inputs besides labor and materials. For example, our methodology is fully consistent with capital adjustment costs. Finally, we also show that our estimation procedure is quantitatively robust to several modifications, including heterogeneous labor within plants, labor adjustment costs, ex-ante specified returns to scale, and alternative definitions of compensation that incorporate benefits.

Since direct measures of employer market power have been scarce, indirect measures based on concentration are commonly-used alternatives. Studying monopsony using concentration is conceptually challenging, as different theoretical frameworks imply either a positive or a negative relationship between a market's extent of *competition* and its level of *concentration* (as recently reiterated, among others, by Syverson, 2019). We find that, though markdowns are increasing with size, the cross-sectional correlation between local labor market concentration and local markdowns does not support the view of concentration being an appropriate proxy for market power — at least within manufacturing. Though aggregate local concentration also displays a decreasing trend since the late 1970s, it fails to mimic the reversal that our series for the aggregate markdown displays since the early 2000s. We conclude that caution must be exercised when using employment concentration to infer levels or trends in employer market power.

Contribution to the literature. Our paper contributes to a recently reinvigorated research agenda on the prevalence and evolution of labor market monopsony in the U.S. economy. We do so in two ways. First, we use comprehensive administrative data for U.S. manufacturers and provide direct estimates of the wedge between an employer's marginal revenue product of labor and its wage. Furthermore, we document a substantial amount of dispersion in markdowns and investigate the source(s) of markdown heterogeneity. We do so by making explicit the relation between markdowns and employer characteristics, such as size, age, productivity and scope. Second, we show how to characterize aggregate markdowns and document its evolution over the past four decades.

Interest in the exercise of market power by firms, and especially large firms, has been recently revived by concerns that increased distortions in U.S. output and input markets have contributed to several secular trends, most notably the decline in the labor share. Documented as a robust macroeconomic phenomenon in both the U.S. (Elsby, Hobijn and Sahin, 2013) and a variety of other countries (Karabarbounis and Neiman, 2013), the decline in the share of income that accrues to labor is the main motivation behind the analysis of industry-level sales concentration in Autor et al. (2017). The authors document a substantial rise in the share of sales accounted for by the largest firms in each industry ("superstar firms") and conclude that such increase is capable of explaining the aggregate decline in labor. Brooks et al. (2019) use techniques analogous to this paper to estimate markdowns in China and India, and similarly conclude that "in the context of developing economies, markdowns substantially lower the labor share". A related literature has documented a contemporaneous increase in markups and suggested that the latter could be a unifying explanation behind many observed secular trends in the U.S. economy, including the decrease in the labor share (De Loecker and Eeckhout, 2018; De Loecker, Eeckhout and Unger, 2018; Eggertsson, Robbins and Wold, 2018). Though our paper does not directly address these questions, it contributes related evidence on the dynamics of the labor share and wages at the micro-level. Specifically, we document substantial variation in plant-level markdowns for the manufacturing sector, both across and within narrowly-defined industries, and illustrate a tight positive relationship of markdowns and size.

Our markdown estimation procedure relies on the so-called "production approach" (De Loecker, 2011) which combines insights from Hall (1988) with production function estimation techniques from the IO literature (Levinsohn and Petrin, 2003; De Loecker and Warzynski, 2012; Ackerberg, Caves and Frazer, 2015). In our baseline measure, we assume that firms take monopsony forces into account by internalizing a finitely elastic labor supply curve, thus reflecting the assumption of an upward-sloping labor supply curve common in many of the current models of monopsony. This includes frameworks based on Burdett and Mortensen (1998), as in Bontemps, Robin and Van den Berg (2001), Manning (2003), Mortensen (2003), and Manning (2011); to the class of additive random utility models as characterized in Chan, Kroft and Mourifie (2019), which include Card et al. (2018) and Lamadon, Mogstad and Setzler (2019); to environments based on monopsonistic competition as in Bhaskar and To (1999), Staiger, Spetz and Phibbs (2010), and Berger, Herkenhoff and Mongey (2019). In particular, Berger, Herkenhoff and Mongey (2019) quantify the welfare cost of labor market power using an oligopsony version of the framework by Atkeson and Burstein (2008). The authors use heterogeneous responses across firms to changes in state-level corporate taxes to identify the key parameters of their structural model. Their model implies that markdowns are increasing in firm size; a prediction that is consistent with our empirical results. Jarosch, Nimczik and Sorkin (2019) also build a model in which firm size is a source of market power. Employer market power arises due to search frictions, as opposed to finitely elastic labor supply curves, thus their framework is not nested within our setup. As a result, we view their work as complementary to ours. Our paper contributes to this literature by proposing a strategy to estimate markdowns that, while compatible with many of the frameworks put forth in previous work, is not tightly linked to a specific micro-foundation but instead is quite general.

In our estimation procedure, we explicitly identify markups and markdowns separately. As a result, we do not confound these two sources of market power. Most previous studies tend to focus on only one source of market power instead and thus overstate the extent of that source's market power. Exceptions to this practice, however, include Dobbelaere and Mairesse (2013), Brooks et al. (2019) and Morlacco (2019) who also exploit the flexibility of material inputs to study monopsony in non-U.S. labor markets and the market for foreign intermediate inputs, respectively.

Standard arguments dating back to Robinson (1933) imply that markdowns are one-to-one with labor supply elasticities. As a result, our markdown estimates also speak to the literature evaluating the elasticity of labor supply. Indeed, most literature refers to monopsony power as a firm's ability to compensate its workers below their marginal revenue product precisely because the elasticity of labor supply *to the firm* is finite (Boal and Ransom, 1997; Manning, 2003; Ashenfelter, Farber and Ransom, 2011). In a perfectly competitive labor market, such elasticity would be infinite and a small drop in the firm-level wage would lead to an exodus of workers towards other employers.

In recent years, several studies have provided empirical evidence of a finite (and small) elasticity of labor supply, though often in very specialized settings and with a somewhat limited scope. These include markets on Veterans Affairs hospitals (Staiger, Spetz and Phibbs, 2010), teachers (Falch, 2010; Ransom and Sims, 2010), nurses (Matsudaira, 2014), online demand platforms (Dube et al., 2018), ride-shares (Caldwell and Oehlsen, 2018), and colleges and universities (Goolsbee and Syverson, 2019). An exception to this very specialized approach is Webber (2015), who uses administrative data for U.S. workers and firms to estimate labor supply elasticities at the employer level. In this paper, on the other hand, we exploit the one-to-one relationship between plant-level markdowns and the elasticity of labor supply to the plant. Our implied elasticity estimates are in line with the literature, with an average of 1.27. This value is not far from Webber's (2015) mean estimate of 1.08 and, as shown by the meta-study by Sokolova and Sorensen (2018), also happens to be the median value when plotting the distribution of elasticity estimates from more than 800 research papers.

Finally, our paper relates to the burgeoning literature on labor market concentration since we evaluate the validity of concentration indexes as proxies for market power. Interest in concentration indexes stems from their ease and breadth of use in both academic research and the practice of antitrust in the U.S. economy. These have been calculated at the national (Autor et al., 2017) and local level (Rossi-Hansberg, Sarte and Trachter, 2018), and show diverging long-run trends. Recent works by Azar, Marinescu and Steinbaum (2017) and Azar et al. (2018) show the negative association between concentration and wages using vacancy data from online sources and argue for an extension of antitrust best practices to the mergers that affect concentration in the labor market. A recent paper by Benmelech, Bergman and Kim (2018) also computes employment concentration and relates it to average wages in U.S. manufacturing. Lipsius (2018) and Rinz (2018) both provide estimates of concentration in firm-level employment from the Longitudinal Business Database and conclude that, though local concentration reduces earnings and increases inequality, observed changes in concentration are unable to explain the rise in income inequality observed in the U.S. economy.

Despite this popular usage, however, it is unclear from a theoretical standpoint whether a market's labor concentration is necessarily positively correlated with its level of competitiveness in the markdown sense (Syverson, 2019). Our paper contributes to this debate by documenting that the covariance between markdowns and employment concentration over time is quite modest. Thus, our results caution against using labor market concentration to infer conclusions on employer market power.

OVERVIEW OF THIS PAPER. Section 2 lays out our estimation procedure and describes the data. Section 3 illustrates our markdown estimates and discusses heterogeneity in markdowns and robustness of our results. Section 4 proposes a novel measure for aggregate markdowns and shows that the time trend in aggregate markdowns is U-shaped, with a minimum in the early 2000s. It concludes with documenting a weak relationship between our estimated aggregate markdown and an index of local employment concentration. Section 5 summarizes the evidence and concludes.

2 Markdown estimation

Our analysis of monopsony in the U.S. labor market is based on markdowns, the percentage gap between a plant's marginal revenue product of labor (MRPL) and the wage it pays its workers. This is a direct measure of employer market power that is easy to compare to the benchmark of perfect competition. In a perfectly competitive labor market, markdowns would be equal to unity. On the other hand, whenever markdowns are larger than unity, the employer compensates workers less than dollar-for-dollar for every unit of revenue produced at the margin.

In this section, we first set out our basic framework and use the optimality conditions from a firm's profit maximization problem to show a one-to-one relationship between the elasticity of the labor supply at the firm-level and markdowns. Then, we use the firm's dual problem (through cost minimization) to set out an estimation strategy in the spirit of Hall (1988) and De Loecker and Warzynski (2012). Using this strategy and detailed administrative data on plants' output and inputs, we retrieve micro-level markdowns in the U.S. manufacturing sector. Interestingly, such an environment also allows for positive product markups, which we estimate alongside markdowns.

2.1 Obtaining markdowns through duality

2.1.1 Profit maximization

Our notion of an individual employer's monopsony power is rooted in the idea that a monopsonistic employer can compensate its workers below their marginal revenue product of labor, a definition of monopsony power popularized by Manning (2003). As said, we refer to this gap, expressed in percentage terms, as a firm's markdown. In the following, we will show that a firm's markdown has an one-to-one relationship with its perceived labor supply elasticity.[2] To do so, consider a firm's profit maximization problem:

$$\max_{\ell \geq 0} R(\ell) - w(\ell)\ell$$

where $R(\ell) \equiv \text{rev}(\ell; \mathbf{x}^*_{-\ell}(\ell))$ is shorthand notation for revenues in which all inputs are evaluated at their optimum with the exception of labor ℓ. For ease of notation, we drop the firm's index for the moment. Given this structure and assuming that the revenue function and wage schedule are differentiable, a firm's optimality condition can be rearranged as:

$$\begin{aligned} R'(\ell^*) &= \left[\frac{w'(\ell^*)\ell^*}{w(\ell^*)} + 1\right] w(\ell^*) \\ &= \left[\varepsilon_S^{-1} + 1\right] \cdot w(\ell^*) \end{aligned} \tag{1}$$

where the firm's perceived (inverse) elasticity of labor supply is defined as: $\varepsilon_S^{-1} \equiv \frac{w'(\ell)\ell}{w(\ell)}\big|_{\ell=\ell^*}$. Therefore, it is sufficient to characterize a firm's labor supply elasticity in order to retrieve its markdown. Hence, we get:

$$\nu \equiv \frac{R'(\ell^*)}{w(\ell^*)} = \varepsilon_S^{-1} + 1 \tag{2}$$

In this conceptual framework, we do not take a specific stance on the sources of monopsony power. Rather, by assuming differentiability, we surmise that monopsony power arises from variation in labor supply elasticities at the firm level. In Appendix D, we show that our setup is quite general and nests a variety of monopsony frameworks, including wage-posting models à la Burdett and Mortensen (1998), additive random utility models (Chan, Kroft and Mourifie, 2019; Card et al., 2018 and Lamadon, Mogstad and Setzler, 2019), and monopsonistic competition models (Bhaskar and To, 1999; Staiger, Spetz and Phibbs, 2010 and Berger, Herkenhoff and Mongey, 2019).

[2] Note that this parallels the intuition behind the Lerner index formula which relates residual demand elasticities with price-cost markups.

2.1.2 Cost minimization

Estimating a firm's perceived elasticity of labor supply in a general setting is challenging. In this section, we use insights from the industrial organization (IO) literature to circumvent this problem. We propose a methodology to retrieve markdowns for U.S. manufacturers that builds on insights from Hall (1988) and De Loecker and Warzynski (2012), sometimes referred to as the "production approach". The key insight is that wedges between output elasticities and revenue shares reflect market power in either output or input markets (or both). Intuitively, the output elasticity of labor captures the gain of an additional unit of labor, whereas labor's share of revenue reflects its cost (normalized by a firm's total revenue). If this wedge is larger than unity, the marginal gain is larger than its costs and the firm must be capturing margins through either markups on its output or markdowns on its inputs.

We now formalize this intuition. Consider a firm's conditional cost minimization problem:

$$\min_{\ell \geq 0} w(\ell) \cdot \ell \ \text{ s.t. } \ F(\ell) \geq Y \tag{3}$$

where, with some abuse of notation, $F(\ell) = F(\ell; \mathbf{x}^*_{-\ell}(\ell))$ denotes a firm's production function in which all inputs are evaluated at their cost-minimizing quantities with the exception of labor. Denoting the associated Lagrangian multiplier by λ, a firm's optimality condition can be written as:

$$\frac{w'(\ell) \cdot \ell}{w(\ell)} + 1 = \lambda \frac{F'(\ell)}{w(\ell)}$$

After some manipulation, we can rewrite the above equation as:

$$\varepsilon_S^{-1} + 1 = \mu^{-1} \cdot \theta_\ell \cdot \alpha_\ell^{-1} \tag{4}$$

where $\alpha_\ell \equiv \frac{w(\ell)\ell}{R(\ell)}\big|_{\ell=\ell^*}$ denotes a firm's labor share of revenues, $\theta_\ell \equiv \frac{F'(\ell)\ell}{F(\ell)}\big|_{\ell=\ell^*}$ is the elasticity of output with respect to labor, and μ is a firm's price-cost markup, i.e. the ratio between its price and marginal cost of production. Recall that markdowns satisfy the equality $\nu = \varepsilon_S^{-1} + 1$. Then, equation (4) shows that the ratio between the output elasticity with respect to labor and labor's revenue share reflects market power in the output market (markups) and/or in the labor market (markdown):

$$\nu \cdot \mu = \frac{\theta_\ell}{\alpha_\ell}$$

When additional data on at least one *flexible* input is available, the same logic as in equation (4) leads us to conclude that the wedge between a flexible input's output elasticity and its revenue share can reflect only markups. In other words, the presence of a flexible production input allows us to separate markdowns from markups. In the remainder of this paper, we follow the IO literature and assume that material inputs (indexed by M) are flexible — i.e., they are not subject to monopsony forces or adjustment costs.[3] This assumption, together with differentiability, is summarized below.

ASSUMPTION 1. Each firm engages in cost minimization and production functions are continuously differentiable. Furthermore, material inputs M are flexible.

Given this assumption, we can then state our main result:

PROPOSITION 1. Let the wage schedule $w(\ell)$ be continuously differentiable and suppose labor is chosen statically, then markdowns satisfy:

$$\nu = \frac{\theta_\ell}{\alpha_\ell}\left(\frac{\theta_M}{\alpha_M}\right)^{-1} \tag{5}$$

Proof. See Appendix A.1. □

As a consequence of equation (5), we can construct markdowns if we have estimates for output elasticities and the revenue shares of labor and materials. We rely on comprehensive administrative data from U.S. manufacturing plants (detailed in section 2.4) in which we can directly observe revenue shares and estimate output elasticities.

Our approach requires only the existence of *at least one* flexible input. Hence, our estimation methodology is completely consistent with adjustment costs for (durable) capital. On the other hand, we do assume that labor adjustment costs are absent in our baseline estimation. Section 3.3 and Appendix C show that, however, incorporating labor adjustment costs results in quantitatively minimal adjustments to our baseline estimates.

2.2 Production function estimation

Proposition 1 indicates that observing output elasticities and revenue shares is sufficient for constructing markdowns. Revenue shares are directly observable in administrative data on U.S. manufacturing plants, although output elasticities need to be estimated. To do so, we rely on "proxy variable" methods to estimate production functions (see Levinsohn and Petrin, 2003; De Loecker and Warzynski, 2012; Ackerberg, Caves and Frazer, 2015). One main advantage of the proxy variable approach is that we can explicitly distinguish between market power in output and labor markets. This is possible as long as a flexible input exists and output elasticities are estimable Assumption 1 postulates that materials are a flexible input. However, the estimation of output elasticities requires additional assumptions. We adopt the standard assumptions of the proxy variable literature, which are informally summarized below. We refer the reader interested in a formal treatment of our estimation procedure, including a detailed set of sufficient conditions, to Appendix A.

[3] While we focus on the economic intuition behind our approach in the present discussion, the formal assumptions required to obtain equation (4) are stated in full detail in Appendix A. In section 3.3 and Appendix C we also show that allowing for labor adjustment costs (convex or non-convex) does not alter our results qualitatively or quantitatively.

ASSUMPTION 2. Each firm has a translog production function for gross output in capital, labor, material inputs and energy. Production function parameters are constant over time and common within an industry group.

ASSUMPTION 3. A firm's productivity is Hicks-neutral and has the Markov property. Innovations to productivity in period $t+1$ are orthogonal to inputs that are chosen in period t. Furthermore, material inputs are monotonic in productivity.

Under the production approach, we do not need to make any assumptions on the underlying sources of market power. Hence, it is not necessary to impose functional forms on residual demand functions or labor supply curves. On the other hand, we do need to make assumptions on a firm's production technology. Assumption 2 specifies a rather flexible functional form for the production function, i.e. translog production. Its flexibility rests on its interpretability as a second-order approximation to *any* arbitrary production function (see De Loecker and Warzynski, 2012). Hence, a translog specification nests and is substantially more general than, for example, a Cobb-Douglas specification.

Assumption 2 allows production parameters to vary across detailed industry groups (i.e., 3-digit NAICS) but imposes that they are constant over time. However, this does not imply that *output elasticities* are constant over time. Indeed, under a translog specification for gross output, output elasticities are allowed to vary across plants with the (time-varying) level of each firm's inputs.[4]

Assumption 3 imposes restrictions on plant- or firm-level productivity (see Levinsohn and Petrin, 2003; De Loecker and Warzynski, 2012). In particular, production function parameters are informed through a set of moment conditions in which inputs chosen in period t are orthogonal to innovations to productivity in period $t+1$. The last statement of assumption 3 guarantees that material inputs are invertible in productivity. This particular condition allows us to estimate production function parameters without directly observing productivity.

While the conditions in assumptions 2 and 3 may appear restrictive at first sight, they nest standard assumptions in canonical models of firm dynamics. In particular, our approach nests Cobb-Douglas technologies with productivity processes that display significant persistence through, for example, an AR(1) specification.

We denote y_{it} as log output whereas $\mathbf{x}_{it} = (k_{it}, \ell_{it}, m_{it}, e_{it})'$ denotes the vector of log inputs (namely, capital, labor, materials and energy) and $\mathbf{z}_{it}$ is the vector instrumenting for the set of endogenous inputs $\mathbf{x}_{it}$. Furthermore, $f(\mathbf{x}_{it}; \boldsymbol{\beta})$ denotes the log transformation of the production function. Given assumptions 2 and 3, we estimate production function parameters $\boldsymbol{\beta} \in \mathbb{R}^Z$ for each production function in a three-step process:

[4] Furthermore, explicitly allowing time-varying production parameters does not greatly alter our conclusions. As a result, this part of Assumption 2 is without much loss of generality.

1. Run a polynomial regression of y_{it} on $\mathbf{x}_{it}$ and a set of controls. Obtain non-parametric estimate of log output φ_{it} free of measurement error.

2. Construct productivity as $\omega_{it}(\boldsymbol{\beta}) = \varphi_{it} - f(\mathbf{x}_{it}; \boldsymbol{\beta})$ and run a polynomial regression of $\omega_{it}(\boldsymbol{\beta})$ on $\omega_{it-1}(\boldsymbol{\beta})$ to obtain productivity shocks $\xi_{it}(\boldsymbol{\beta})$.

3. Estimate production function parameters $\boldsymbol{\beta}$ through the GMM system induced by the moment conditions $\mathbb{E}_{t-1}\left(\xi_{it}(\boldsymbol{\beta}) \cdot \mathbf{z}_{it}\right) = \mathbf{0}_{Z\times 1}$.

Once the production function parameters $\boldsymbol{\beta}$ are obtained, it is straightforward to calculate output elasticities. Under a Cobb-Douglas specification, for example, the parameters $\boldsymbol{\beta}$ are equal to output elasticities. However, under our translog setup, output elasticities are a linear function of the inputs $\mathbf{x}_{it}$, with coefficients that depend on $\boldsymbol{\beta}$. A complete description on the construction of output elasticities under translog technologies is found in Appendix A.

2.3 Discussion

We regard the generality of our approach as one of this paper's major attractive points. Indeed, an advantage of the production approach is that we do not need to make any assumptions on the sources of market power in order to quantify markdowns. In particular, we do not take a stance on the market structure for labor or the form of labor supply curves that firms face — a distinguishing feature of this paper with respect to the fully-developed structural efforts of, for example, Card et al. (2018), Berger, Herkenhoff and Mongey (2019), Chan, Kroft and Mourifie (2019), and Lamadon, Mogstad and Setzler (2019). Furthermore, we do not need to make any assumptions regarding the flexibility of other inputs besides materials. Our approach is valid as long as firms are subject to some finitely elastic labor supply curve and material inputs are flexible. In Appendix D, we emphasize the richness of our methodology by describing a broad range of monopsony models that our approach permits.

The production approach, as popularized by De Loecker and Warzynski (2012), comes with many advantages but is not free of criticism. One of the key identifying assumptions ris the requirement for at least one flexible input. Pinpointing such an input is difficult in most publicly-available data sets, since most inputs are not registered separately but rather classified into groups following accounting standards.[5] While some of the discussion on measuring market power has revolved on what constitutes a flexible input (e.g., Traina, 2018), we follow the IO literature and assume that material inputs are flexible (Basu, 1995; De Loecker and Warzynski, 2012).

Despite this standard IO assumption, there is some evidence of monopsony in the market for material inputs. For instance, Morlacco (2019) uses transaction-level data from French manufacturers to present evidence of

[5] Recent studies that identify markups typically rely on the Compustat database, in which variable inputs are often identified with "cost of goods sold" (COGS) — which includes material inputs as well as variable and fixed labor — or "selling, general, and administrative expenses" (SGA). We circumvent whether COGS or SGA is a more appropriate flexible input and directly use material inputs instead. We can do so because expenditures on capital, labor, material and energy inputs are *separately* observed in our data.

market power in imported intermediate inputs, under the identifying assumption that domestically-sourced intermediate inputs are instead perfectly competitive. If material inputs are subject to monopsony, then the ratio $\frac{\theta_\ell}{\alpha_\ell}\left(\frac{\theta_M}{\alpha_M}\right)^{-1}$ in equation (5) would reflect the markdown for labor *relative to the markdown for materials*, say ν_ℓ/ν_M. Therefore, in the presence of market power for materials, we would obtain $\nu_M > 1$ and our estimates for labor markdowns would actually be underestimated. In other words, our conclusions on the extent of monopsony in the U.S. labor market would only be reinforced, as labor markdowns would be even greater than what we estimate.

A plausible alternative for the flexible input is energy inputs, as advocated by Kim (2017). He maintains that monopsony power through buyer-supplier networks may be a concern for materials, while energy inputs are less prone to monopsony forces as prices for energy tend to be regulated. On the other hand, Davis et al. (2013) provide robust evidence against the hypothesis that energy inputs are not subject to monopsony. They find that plant-level differences within manufacturing industries in energy purchases account for a substantial fraction of overall price dispersion.[6] Furthermore, price gaps between larger and smaller purchases are large, even when controlling for plant location and/or electric utility provider fixed effects. This seems to contradict the "no monopsony" condition for energy and suggests that material inputs may be a better choice. Finally, material inputs represent a much larger share of revenues in manufacturing than energy inputs do. This implies that measurement error is of lesser concern for material inputs compared to energy. We view these results as compelling evidence in favor of materials as flexible inputs.

A remaining potential point of concern lies in whether the imposed "proxy structure" achieves point identification. In particular, Gandhi, Navarro and Rivers (2017) show that the assumptions of the proxy variable method are insufficient to point-identify production function parameters, and that additional sources of variation in the demand for flexible inputs are required. Recently, Flynn, Gandhi and Traina (2019) have shown that point identification can be restored whenever the returns to scale of the production function is known. They suggest that constant returns to scale is a useful benchmark and this additional assumption performs "remarkably well" in their Monte Carlo simulations. Hence, we subject our results to their methodology and confirm they are robust to this additional assumption. A more detailed discussion can be found in Appendix A, where we also lay out the exact moment conditions for markdown estimation under constant returns to scale. Section 3.3 reports estimated markdowns under different methodologies and highlights how our conclusions are unchanged whenever we impose the conditions in Flynn, Gandhi and Traina (2019).

2.4 Data: Censuses and Annual Surveys of Manufactures

We use two administrative datasets for the estimation of markdowns: the Census of Manufactures (CM) and the Annual Survey of Manufactures (ASM), both from the U.S. Census Bureau. The Census of Manufacturing is a quinquennial survey that covers the universe of manufacturing establishments in years ending in "2" and "7". The main advantage of using the CM is that it contains establishment-level data on revenues and

[6] According to Davis et al. (2013), at least one-third of the cross-sectional dispersion in log electricity prices is due to variation between purchase deciles. This fraction was as high as 75 percent in 1963 but dropped to 30 percent by 1978.

inputs, the two necessary ingredients for production function estimation. All our measures of output (revenues) and inputs, such as capital, labor, material and energy inputs, are constructed with the use of deflators from the NBER-CES Manufacturing Database and follow standard procedures that are described extensively in, for example, Syverson (2004a) and Kehrig (2015).

In order to construct markdowns for non-Census years, we use the Annual Survey of Manufactures (ASM). The ASM contains a representative sample of manufacturing plants that rotates in years ending in "4" and "9". While large plants are sampled with probability near unity, small plants are less frequently sampled.[7] We use Census-provided sampling weights to ensure that our estimates are representative of the whole manufacturing sector. Our main results in the following subsections are based on a non-balanced panel for manufacturing plants in years 1976–2014. To avoid artificial spikes in Census years, unless noted otherwise, we keep only those plants that are in the rotating sample of the ASM in Census years.

3 Markdowns in U.S. manufacturing

3.1 Cross-sectional distribution

The results of our estimation procedure are displayed in table I and paint a clear picture: markdowns are sizable and considerably larger than unity. The average establishment charges a markdown of 1.53 — that is, a plant's marginal revenue product of labor is on average 53 percent higher than the wage it pays its workers. To put these numbers in perspective, a markdown of 1.53 implies that a worker receives about 65 cents on the marginal dollar generated. Furthermore, we find that a large number of manufacturing plants exert labor market power. In fact, half of the manufacturing plants charge a markdown that is equal to or larger than 1.364 (73 cents on the dollar). Despite the fact that our estimated markdowns are significantly larger than one, it is worth noting that these levels are largely in line with those from previous studies (see Manning, 2003; Sokolova and Sorensen, 2018). When we compare our markdown estimates with the meta-analysis on labor supply elasticities by Sokolova and Sorensen (2018), our results are actually in the lower half of this literature. We conclude that the data support the hypothesis that the average (or even median) manufacturing plant operates in a monopsonistic environment.

Moreover, a large part of the variation in markdowns originates across plants *within* the same industry. The average *within-industry* interquartile range (standard deviation) of markdowns is 61.6 (60.4) percent. This fact suggests that heterogeneity in markdowns is likely related to idiosyncratic factors, such as plant-level productivity differences or specific human capital, rather than industry-wide characteristics, such as legacy structure, institutional agreements, or industry regulations.[8]

This heterogeneity remains robust when looking at within-industry standard deviations of markdowns. As a recent literature has emphasized that the welfare cost of market power distortions can be considerable

[7] Plant size is determined by Census in terms of revenues and/or employment.

[8] This pattern accords with the dispersion in revenue-based total factor productivity documented by Syverson (2004b).

Table I: Estimated plant-level markdowns in U.S. manufacturing: markdowns are sizable and considerably larger than unity. The average manufacturing plant operates in a monopsonistic environment.[a]

INDUSTRY GROUP	Median	Mean	IQR_{75-25}	SD
Petroleum Refining	2.391	2.547	1.828	1.267
Computer and Electronics	2.296	2.558	1.227	1.075
Plastics and Rubber	1.812	1.906	0.582	0.584
Food and Kindred Products	1.761	1.913	0.872	0.823
Paper and Allied Products	1.695	1.795	0.573	0.625
Chemicals	1.623	1.817	0.941	0.870
Lumber	1.540	1.623	0.467	0.522
Primary Metals	1.450	1.503	0.506	0.479
Electrical Machinery	1.317	1.416	0.519	0.513
Motor Vehicles	1.368	1.422	0.376	0.432
Printing and Publishing	1.345	1.495	0.454	0.632
Fabricated Metal Products	1.257	1.313	0.339	0.360
Non-electrical Machinery	1.246	1.317	0.532	0.454
Miscellaneous Manufacturing	1.208	1.254	0.348	0.358
Textile Mill Products	1.208	1.266	0.412	0.454
Furniture and Fixtures	1.150	1.167	0.320	0.358
Non-metallic Minerals	1.139	1.217	0.372	0.522
Apparel and Leather	1.035	1.146	0.413	0.539
Whole sample	**1.364**	**1.530**	**0.618**	**0.708**
Sample size	$1.393 \cdot 10^6$			

[a] Markdowns are estimated under the assumption of a translog specification for gross output. Each industry group in manufacturing corresponds to the manufacturing categorization of the BEA which approximately follows a 3-digit NAICS specification. Source: authors' calculations from ASM/CM data in 1976–2014.

(Baqaee and Farhi, 2018; Berger, Herkenhoff and Mongey, 2019; Edmond, Midrigan and Xu, 2019), it is important to focus on understanding the determinants of markdown variation.

3.2 Heterogeneity in markdowns

To better understand the main source of variation in markdowns, we first decompose markdowns into their components according to equation (4). Micro-level markdowns are additively separable (in natural logs) in the following terms:

$$\ln(\nu) = \ln(\theta_\ell) - \ln(\alpha_\ell) - \ln(\mu) \tag{6}$$

Recall that θ_ℓ is the elasticity of output with respect to labor, α_ℓ is labor's share of revenue, and μ is the product markup. We can then apply the following variance decomposition:

$$\begin{aligned} V(\ln(\nu)) &= V(\ln(\theta_\ell)) + V(\ln(\alpha_\ell)) + V(\ln(\mu)) \\ &\quad - 2 \cdot [\text{cov}(\ln(\theta_\ell), \ln(\alpha_\ell)) - \text{cov}(\ln(\alpha_\ell), \ln(\mu)) + \text{cov}(\ln(\theta_\ell), \ln(\mu))] \end{aligned} \tag{7}$$

In table II we document the contribution of each component. The variation in markdowns is largely accounted for by the variation in output elasticities θ_ℓ and labor shares α_ℓ. Importantly however, the covariance terms are non-negligible. In fact, the covariance term between output elasticities and labor shares is of a similar magnitude to the previous two mentioned variances.

Table II: Variance in plant-level markdowns is accounted for by the variances and covariances of output elasticities θ_ℓ and labor shares α_ℓ. Variance in markups is quantitatively small.[b]

		Variance	Relative contribution
Markdown	ν	0.1696	1.000
Elasticity	θ_ℓ	0.3149	1.857
Labor share	α_ℓ	0.3813	2.248
Markup	μ	0.0276	0.1627
		Covariance	Relative contribution
	θ_ℓ, α_ℓ	0.2804	1.653
	θ_ℓ, μ	−0.00601	−0.0354
	α_ℓ, μ	−0.00271	−0.0160

[b] Variance decomposition of plant-level markdowns as based on equation (7). Source: authors' calculations from ASM/CM data in 1976–2014.

On the other hand, variations in markups play a quantitatively small role for markdown variation.[9] Hence, our results imply that the main determinants of markdown variation are different from those levers that drive variation in markups.

Size, age, and productivity. We proceed to investigate the source(s) of markdown variation by focusing on idiosyncratic factors. In particular, we look at the relationship between markdowns and establishment size. A recent literature has emphasized the welfare costs of markups and markdowns that vary through size alone. For instance, Edmond, Midrigan and Xu (2019) focus on markup distortions induced by size as modeled through a Kimball (1995) aggregator. The authors show that the welfare costs of markups are large in such an environment. Similarly, Berger, Herkenhoff and Mongey (2019) allow for size-varying markdowns in the spirit of Atkeson and Burstein (2008) and find considerable welfare losses from firms' labor market power. Therefore, it is natural to ask whether size can account for a substantial amount of variation in our markdown estimates.

As mentioned by Haltiwanger, Jarmin and Miranda (2013), however, it is important to control for age while assessing size effects because the two are heavily correlated and could thus confound each other. Therefore, we run a set of non-parametric regressions in the spirit of Haltiwanger, Jarmin and Miranda (2013) to fully capture the heterogeneity of markdowns in size and age. We report two sets of results. First, we focus on the size-markdown relationship without controlling for age. Then, we document how our size coefficients vary

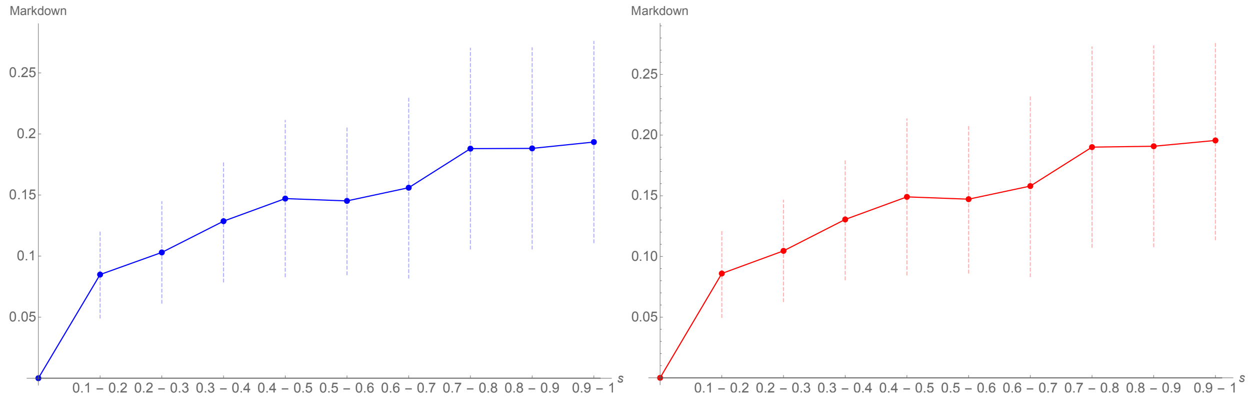
[9] Although the focus of this paper is on the estimation of markdowns, we acknowledge that a more complete treatment of the relationship between plant-level markups and markdowns is worthy of future research.

when introducing age controls. Our set of regressions are of the following form:

$$\ln(\nu_{it}) = \beta_0 + \sum_{d=1}^{\mathcal{S}} \beta_d^{\text{size}} \cdot \mathbf{1}_{s_{it} \in S_d} + \sum_{d=1}^{\mathcal{A}} \beta_d^{\text{age}} \cdot \mathbf{1}_{\text{age}_{it} \in A_d} + \mathbf{X}_{it}' \gamma + \varepsilon_{it} \tag{8}$$

where $\mathbf{X}_{it}$ contains a full set of industry and year fixed effects.[10] In our baseline regression, size dummies are created over a grid of $\mathcal{S} = 10$ equally spaced bins in a plant's employment share (of its local labor market).[11] Furthermore, we categorize age in $\mathcal{A}$ groups, defined similarly as in Haltiwanger, Jarmin and Miranda (2013).[12] The results are depicted in figure 1 and display a clear picture: markdowns are increasing in size. The coefficients on the size dummies are monotonically increasing and statistically significant at the one percent level. On average, plant size can induce markdown variations on the order of roughly 20 percent.

Figure 1: Markdowns are increasing in size. Plant size can induce markdown variations in the order of roughly 20 percent.



Non-parametric size-markdown regressions (employment-weighted). Regression coefficients without (with) age controls are depicted in red (blue). To avoid collinearity issues, we follow Haltiwanger, Jarmin and Miranda (2013) and apply the normalization $\beta_1^{\text{size}} = 0$ (local employment share between 0 and 10 percent). Hence, size coefficients should be interpreted as deviations relative from this baseline. Standard errors are clustered at the industry level. Source: authors' own calculations from ASM/CM data in 1976–2014.

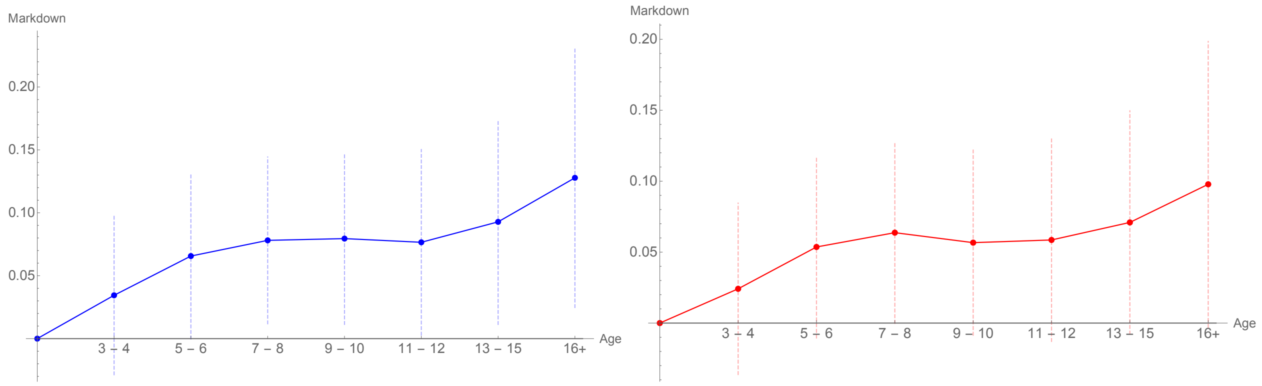
The results for age are similar but portray a less clear-cut picture. Whenever we do not include size controls, markdowns increase monotonically with plant age. However, this positive relationship nearly disappears when we include size controls. Furthermore, the coefficients on the age dummies are much noisier. These results can be found in figure 2. As a result, we are hesitant in establishing a robust markdown-age relationship.

[10] By construction, our size regressions that do not control for age have $\beta_d^{\text{age}} = 0$ for all $d \in \{1, \ldots, \mathcal{A}\}$.

[11] To avoid the majority of our observations being bunched in groups with low labor market shares, we apply employment weights, following Haltiwanger, Jarmin and Miranda (2013), such that the groups correspond to deciles of workers. Our results are barely affected when we do not use employment weights.

[12] Even though the ASM reflects a representative panel of manufacturing plants, biases can occur when assessing a plant's size or age based on information from the ASM/CM alone. To avoid these problems, we take a plant's employment share and age from the Longitudinal Business Database, which contains the universe of employers, and can be merged to ASm/CM at the establishment-year level.

Figure 2: Markdowns increase monotonically with plant age but this result is not robust to size controls.



Non-parametric age-markdown regressions (employment-weighted). Regression coefficients without (with) size controls are depicted in red (blue). To avoid collinearity issues, we follow Haltiwanger, Jarmin and Miranda (2013) and apply the normalization $\beta_1^{\text{age}} = 0$ (age 0 or 1). Hence, age coefficients should be interpreted as deviations relative from this baseline. Standard errors are clustered at the industry level. Source: authors' own calculations from ASM/CM data in 1976–2014.

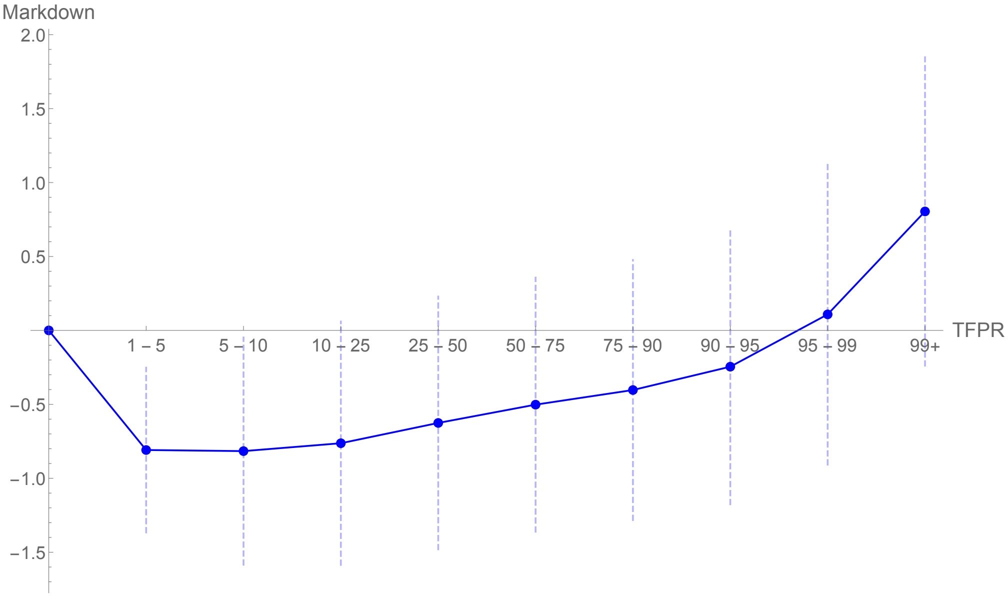
We also investigate the relationship between markdowns and productivity, following the connection made by the rent-sharing literature between wages and profits or sales per worker. Christofides and Oswald (1992) find a robust relationship between industry profits and firm-level wages, while Van Reenen (1996) documents that innovative firms tend to pay their workers higher wages. Recently, Card, Devicienti and Maida (2014) estimate an elasticity of wages to (economic) rents of approximately 4 percent. In general, the literature (see Abowd, Kramarz and Margolis, 1999) finds a positive relationship between wages and a firm's productivity, partially due to positive sorting and partially accounted for by a positive correlation between workers' bargaining power and firms' profitability measures. Our results relate to the latter theme as we correlate plant-level productivity not simply to the average wage but rather to its markdown — the ratio between the marginal revenue product and the wage.

Unlike the results for size and age, we find no monotonic relationship between a plant's markdown and productivity, as depicted in figure 3. Since we do not observe quantities, we proxy physical productivity (TFPQ) by revenue productivity (TFPR).[13]

The data suggest a weak U-shaped relationship between markdowns at productivity, since markdowns are increasing in TFPR only after the 5th percentile in the TFPR distribution. However, the coefficients on the productivity dummies are noisily estimated, and the majority of these estimates are not significantly different from zero at the 5 percent level.

[13] Even though there are differences between TFPR and TFPQ, Foster, Haltiwanger and Syverson (2008) show that TFPQ and TFPR are highly correlated with each other in a subsample of manufacturing plants for which both measures of productivity can be constructed.

Figure 3: There is a weak U-shaped relationship between plant-level markdowns and TFPR.



Non-parametric productivity-markdown regressions (employment-weighted). To avoid collinearity issues, we follow Haltiwanger, Jarmin and Miranda (2013) and apply the normalization $\beta_1^{\text{TFPR}} = 0$ (lower percentile of the TFPR distribution). Hence, productivity coefficients should be interpreted as deviations relative from this baseline. Standard errors are clustered at the industry level. Source: authors' own calculations from ASM/CM data in 1976–2014.

Scope and high-tech status. In the following, we document whether markdowns vary for plants belonging to firms with multiple establishments or with a wide industrial or geographical scope. We create a binary variable that equals one whenever a plant is owned by a firm that has at least two active establishments. Similarly, we also create a binary variable that equals one whenever a plant belongs to a firm with establishments in two or more different 6-digit NAICS industries. If markdowns are higher for these type of plants, we denote this as an "industry scope" premium. Its geographical counterpart is analogously defined based on 5-digit FIPS counties and denoted as a "geographic scope" premium.

Table III shows that plants owned by multi-unit firms charge higher markdowns. Furthermore, we find that industrial and geographical scope premia are positive and average around 25 percent (see table III). These results continue to hold whenever we control for firm size.

Lastly, we investigate whether plants in the high-tech sector have higher markdowns.[14] High-tech firms play a disproportionate role in aggregate employment and productivity growth (see Decker et al., 2016), thus it is interesting to know whether they charge higher markdowns on their labor. We find, however, that high-tech plants feature *lower* average markdowns, which is consistent with the results on more innovative firms in Van Reenen (1996).

[14] We follow the definition for high-tech sectors of Decker et al. (2016). For manufacturing, these include the 4-digit NAICS industries 3254 (pharamceuticals), 3341 (computers), 3342 (communications equipment), 3344 (semiconductors and electronic components), 3345 (precision and control instruments) and 3364 (aerospace).

Table III: Plants belonging to multi-unit firms or firms active in more than one sector/location have higher markdowns. Plants in hi-tech sectors have lower markdowns.[c]

Dependent variable: LOG MARKDOWNS				
	MULTI-UNIT	INDUSTRIAL	GEOGRAPHICAL	HIGH-TECH
Premium	0.2514 (0.04236)	0.2543 (0.04173)	0.2558 (0.04247)	-0.09054 (0.1081)
Observations (in millions)	1.393	1.393	1.393	1.393
R^2	0.2668	0.2696	0.2697	0.2511

[c] Markdowns are estimated under the assumption of a translog specification for gross output. Each industry group in manufacturing corresponds to the manufacturing categorization of the BEA which approximately follows a 3-digit NAICS specification. Industrial and geographical scope refers to a plant that is owned by a firm that is active in multiple 6-digit NAICS and 5-digit FIPS counties, respectively. A plant is considered to be part of the high-tech sector if its 4-digit NAICS code belongs to the list of high-tech industries as defined in Decker et al. (2016). Standard errors are clustered at the industry level and denoted between parentheses. Source: authors' calculations from ASM/CM data in 1976–2014.

HETEROGENEOUS LABOR. Our baseline estimates allow for different types of labor across plants, but we implicitly assumed that labor was homogeneous within plants. To capture potential heterogeneity in markdowns for workers of various skills, we exploit the fact that the ASM and CM break down the wage bill into components of production and non-production workers.. The Census Bureau defines the former as "workers engaged in fabricating, processing, assembling, inspecting, receiving, packing, warehousing, shipping (but not delivering), maintenance, repair, janitorial, guard services, product development, auxiliary production for plant's own use, record keeping, and other closely associated services". This includes line-supervisors but not managerial and administrative positions. This breakdown allows us to estimate production functions with two different types of labor.

We obtain markdown estimates for production and non-production workers for each plant-year observation. Table IV illustrates that allowing for labor heterogeneity does not greatly affect our original estimates. We conclude that our baseline results are not driven by a particular group of workers. In particular, markdowns for non-production workers correspond closely to our baseline estimates whereas the cross-industry heterogeneity of markdowns for production workers is slightly higher than the baseline. Somewhat surprisingly though, markdowns for one group are not systematically higher or lower than for the other.

3.3 Robustness

EX-ANTE SPECIFIED RETURNS TO SCALE. Gandhi, Navarro and Rivers (2017) have shown that the proxy variable methodology does not point-identify production function parameters. This is a potential concern since our identification strategy rests on properly estimating output elasticities (which are a function of production function parameters). Recently, Flynn, Gandhi and Traina (2019) have shown, however, that identification can be restored if returns to scale of the production function are fixed. We follow their approach and re-estimate our markdowns but impose constant returns to scale. The results, displayed under the column

Table IV: Markdowns for both production and non-production workers are close to the baseline estimates in table I. Markdowns for one group are not systematically higher than for the other.^{d}

	Non-production		*Production*	
INDUSTRY GROUP	Mean	Median	Mean	Median
Food and Kindred Products	2.395	2.174	2.014	1.848
Textile Mill Products	1.924	1.736	1.460	1.403
Apparel and Leather	1.311	1.216	1.186	1.122
Lumber	1.660	1.553	1.707	1.620
Furniture and Fixtures	1.372	1.310	1.199	1.138
Paper and Allied Products	1.232	1.125	2.150	2.049
Printing and Publishing	2.021	1.896	1.243	1.142
Chemicals	1.599	1.400	2.473	2.146
Petroleum Refining	2.682	2.356	2.254	1.804
Plastics and Rubber	1.398	1.317	1.802	1.713
Non-metallic Minerals	1.299	1.204	1.628	1.504
Primary Metals	1.824	1.760	1.416	1.339
Fabricated Metal Products	1.474	1.384	1.53	1.422
Non-electrical Machinery	1.539	1.359	5.018	4.530
Electrical Machinery	1.383	1.311	1.667	1.526
Motor Vehicles	1.450	1.411	1.523	1.439
Computer and Electronics	2.620	2.436	3.383	2.954
Miscellaneous Manufacturing	1.532	1.456	1.344	1.258
Baseline	1.530	1.364		
Sample size	$1.393 \cdot 10^6$			

[d] Markdowns are estimated under the assumption of a translog specification for gross output. Each industry group in manufacturing corresponds to the manufacturing categorization of the BEA which approximately follows a 3-digit NAICS specification. The summary statistics under "Non-production" ("Production") reflect markdowns applied to non-production (production) workers. Source: authors' calculations from ASM/CM data in 1976–2014.

"CRS" in table V, do not change substantially from the baseline, even in a quantitative sense, corroborating the notion that our strategy yields reliable estimates of monopsony power in U.S. manufacturing.[15]

LABOR ADJUSTMENT COSTS. In our baseline specification, we assumed away the role of labor adjustment costs. This stems from Assumption 1, in which labor is chosen in a static fashion, ruling out costs from adjustment. Adjustment costs, however, also can potentially drive a wedge between the output elasticity of labor and its revenue share, possibly contaminating our markdown estimates as an expression of monopsony power. In a quantitative assessment of such bias, however, we find that the impact of labor adjustment costs on our estimates is minimal.

To show that adjustment costs result only in quantitatively trivial adjustments to our baseline estimates, we proceed in two steps. First, we show that, when labor is subject to costly convex adjustments, the wedge between the marginal revenue product of labor and wages reflects both monopsony power and adjustment

[15] Our results also are not sensitive to choosing different levels of returns to scale.

Table V: Labor markets in U.S. manufacturing are far from the perfectly competitive ideal and this result is robust to a variety of specifications: ex-ante specified returns to scale (CRS), adjustment costs (Biennial), using wages and benefits as a measure of compensation (Benefits).[e]

INDUSTRY GROUP	Baseline	CRS	Biennial	Benefits
Food and Kindred Products	1.761	1.475	1.871	1.276
Textile Mill Products	1.208	1.389	3.852	1.128
Apparel and Leather	1.035	0.663	1.074	1.024
Lumber	1.540	1.746	1.508	1.223
Furniture and Fixtures	1.150	1.831	1.122	1.038
Paper and Allied Products	1.695	1.669	1.699	1.431
Printing and Publishing	1.345	0.954	1.344	1.263
Chemicals	1.623	1.765	1.671	1.429
Petroleum Refining	2.391	2.826	2.131	3.463
Plastics and Rubber	1.812	1.424	1.200	1.207
Non-metallic Minerals	1.139	1.296	1.289	1.147
Primary Metals	1.450	1.712	1.477	1.440
Fabricated Metal Products	1.257	1.684	1.368	1.148
Non-electrical Machinery	1.246	1.489	1.151	1.068
Electrical Machinery	1.317	1.338	1.184	1.193
Motor Vehicles	1.368	1.663	1.268	1.078
Computer and Electronics	2.296	2.786	2.320	1.669
Miscellaneous Manufacturing	1.208	2.468	1.208	1.114

[e] Markdowns are estimated under the assumption of a translog specification for gross output. For each robustness specification, we report the median of each industry group. Under the column "CRS", we display estimates under the additional assumption of constant returns to scale to deal with identification concerns. Results from estimating markdowns using biennial data to capture non-convex adjustment costs are displayed under the column "Biennial." Results from including benefits in the measure of labor compensation (available only from 2002 forward) are displayed under the column "Benefits." Each industry group in manufacturing corresponds to the manufacturing categorization of the BEA which approximately follows a 3-digit NAICS specification. Source: authors' calculations from ASM/CM data in 1976–2014.

costs. In particular, we show that $\frac{R'(\ell^*)}{w(\ell^*)} = (\varepsilon_S^{-1} + 1) + \mathcal{A}$, where $\mathcal{A}$ would equal zero in the absence of labor adjustment costs. Second, we derive an explicit correction term when labor adjustment costs are quadratic, as commonly suggested in the literature (for example, Hall (2004) and Cooper et al. (2007)). This term depends on a plant's growth in labor and its wage bill, and a parameter governing the magnitude of adjustment costs. When we calibrate the correction term over a varied range of labor adjustments and parameters drawn from the literature, we find that the resulting "corrected" estimates of markdowns are not far from baseline. In particular, the most conservative estimate results in average markdowns being adjusted by no more than 3.15 percent, quite small relative to the average markdown of 1.53. A formal proposition and detailed illustration of our quantitative exercise can be found in Appendix C.

As an additional robustness test for the presence of labor adjustment costs, we re-estimate our markdowns biennially. This is to ensure that our estimates for markdowns are not tainted by fixed or *non*-convex adjustment costs for labor. Conceptually, non-convex adjustment costs that may affect our estimates at the annual frequency are much less likely to do so at the biennial frequency, especially since the majority of plants demonstrate changes in their employment levels every year. Results from this biennial estimation, as illustrated in table V under the column "Biennial", are again similar to baseline, buttressing our confidence in the robustness of our estimated markdowns to general forms of labor adjustment costs.

BENEFITS. Our baseline measure of "wages", used in the results of tables I and IV covers a broad range of compensation, including base salaries and wages, bonuses and incentive pay, and stock grants and options. However, it does not include benefits. Consequently, it is possible that we overestimate employers' labor market power to the extent that health and pension benefits are a significant source of overall labor compensation and correlated with components of markdowns as described in equation (4). We thus re-estimate markdowns at the micro-level including benefits in our compensation measure. Benefits (mainly health and pension) are available only from 2002 onward, which results in a smaller sample than our baseline estimates.[16] Nonetheless, the results — displayed in the last column of table V — are again consistent with our narrative. The distribution of markdowns does not change substantially when we include benefits in labor compensation. Hence, we reinforce our main conclusion that labor markets in U.S. manufacturing are far from perfectly competitive, as this result is robust to a variety of adjustments.

4 Secular trends in aggregate markdowns

4.1 Aggregation

So far, we have mainly focused on markdown dispersion in the cross section and concluded that (i) the average manufacturing plant operates in a monopsonistic environment and (ii) the degree of monopsony power, as measured by plant-level markdowns, varies substantially — even within the same industry — with observables such as size and scope. While an increase in (labor) market power is consistent with several observed secular trends in the U.S. economy, robust evidence for this observation is still absent and "empirical gaps still need to be closed" (Syverson, 2019). In the following, we fill some of these gaps by investigating time trends in *aggregate* markdowns to see whether the degree of labor market monopsony in the U.S. manufacturing sector has increased over time.

Even though we have estimates for markdowns at the plant level, aggregation is not straightforward. Previous contributions have relied on weighted averages based on sales (De Loecker, Eeckhout and Unger, 2018) or employment (Rossi-Hansberg, Sarte and Trachter, 2018), but it is unclear in which context and for which questions it is appropriate to use these particular weights for markdown aggregation. To sidestep these concerns, we propose a measure of aggregate markdowns that is (1) theoretically consistent with aggregate wedges, in the spirit of Edmond, Midrigan and Xu (2019), and (2) also takes the local nature of labor markets into account.

[16] Appendix A.5 provides full details on which benefits are included in the data.

We argue that a measure for aggregate markdowns needs to satisfy these two requirements. First, consistency with aggregate wedges is natural since micro-level markdowns are based on micro-level wedges.[17] Similar approaches have been adopted by, for example, Hsieh and Klenow (2009), and Itskhoki and Moll (2019), who define aggregate productivity as a function of micro-level productivities. Furthermore, we do not have to impose a specific structure for labor or output markets in order to achieve consistency with aggregate wedges. As a result, our measure for the aggregate markdown is then consistent with a variety of monopsony models.

Second, our measure for the aggregate markdown explicitly takes the local nature of labor markets into account. A large body of evidence has shown that labor markets are "local" because workers find it costly to search for jobs and work in locations that are far from where they reside. For instance, Manning and Petrongolo (2017) estimate that the attractiveness of jobs for applicants decays sharply with distance, while Marinescu and Rathelot (2018) find that job seekers are 35 percent less likely to apply to a job that is just 10 miles away from their zip code of residence. At the same time, the U.S. economy is characterized by strong migration responses to adverse local and individual shocks (Kennan and Walker, 2011; Amior and Manning, 2018), alongside substantial occupational and industry switching rates (especially after layoff, as documented by Huckfeldt, 2017 and Macaluso, 2017). Based on all the evidence taken together, we characterize a *local* labor market as a sector-location pair. Our main results employs 3-digit NAICS codes and 5-digits FIPS counties, resulting in a total of more than 200 distinct sectors and over 3,000 locations. In what follows, we denote sectors by j and locations by l.[18]

We define the aggregate markup $\mathcal{M}_{jlt}$ in a labor market (j, l) as the wedge between the aggregate output elasticity of some flexible input and its revenue share.[19] Furthermore, we construct the aggregate markdown $\mathcal{V}_{jlt}$ as that part of the wedge between the aggregate output elasticity of labor and the labor share that is not accounted for by markups. By construction, the following identities hold at the market level:

$$\frac{\theta^L_{jlt}}{\alpha^L_{jlt}} = \mathcal{M}_{jlt} \cdot \mathcal{V}_{jlt} \tag{9}$$

$$\frac{\theta^M_{jlt}}{\alpha^M_{jlt}} = \mathcal{M}_{jlt} \tag{10}$$

where θ^L_{jlt} is the aggregate output elasticity of labor and α^L_{jlt} denotes the labor share in some labor market. These objects are defined analogously for material inputs. We say that any measure for the aggregate markup $\mathcal{M}_{jlt}$ and markdown $\mathcal{V}_{jlt}$, that is based on micro-level markups and markdowns, are consistent with aggregate wedges whenever $\mathcal{M}_{jlt}$ and $\mathcal{V}_{jlt}$ satisfy equations (9) and (10).

[17] Aggregate wedges are consistent with gaps that a fictional representative firm would face. This is the interpretation adopted in, for example, Cole and Ohanian (2002), Gali et al. (2007), and Karabarbounis (2014). In particular, the aggregate wedge that defines the aggregate markdown in our setup is part of the MPN-real wage gap in Karabarbounis (2014).

[18] We thank Jan Eeckhout for his suggestion to explore aggregate markdowns while thinking of the local nature of labor markets.

[19] Edmond, Midrigan and Xu (2019) adhere to a similar definition, but assume that labor is fully flexible instead.

Then, we can show the following:

PROPOSITION 2. Let assumption 1 hold, firm-level wage schedules be differentiable and each firm choose labor statically. Then, the **aggregate markdown** and **aggregate markup** for a local labor market (j, l) are consistent with aggregate wedges whenever they are equal to:

$$\mathcal{V}_{jlt} = \frac{\left(\sum_{i \in F_t(j,l)} s_{it} \cdot \frac{\theta^L_{it}}{\theta^L_{jlt}} \cdot (\nu_{it}\mu_{it})^{-1}\right)^{-1}}{\left(\sum_{i \in F_t(j,l)} s_{it} \cdot \frac{\theta^M_{it}}{\theta^M_{jlt}} \cdot \mu_{it}^{-1}\right)^{-1}} \tag{11}$$

$$\mathcal{M}_{jlt} = \left(\sum_{i \in F_t(j,l)} s_{it} \cdot \frac{\theta^M_{it}}{\theta^M_{jlt}} \cdot \mu_{it}^{-1}\right)^{-1} \tag{12}$$

where s_{it} are sales weights, i.e. $s_{it} = \frac{p_{it}y_{it}}{P_{j\ell t}Y_{jlt}}$.

Proof. See Appendix B.1. □

Whenever the market for material inputs is perfectly competitive, we can use a similar insight to the one used in proposition 1. Specifically, proposition 1 states that firm-level markups are equal to the ratio between the output elasticity for materials and their revenue share. If we define the *aggregate* markup to be equal to the ratio between the *aggregate* output elasticity for materials and their *aggregate* revenue share, then the aggregate markup is a weighted harmonic average of firm-level markups, i.e. $\mathcal{M}_{jlt} = \left(\sum_{i \in F_t(j,l)} s_{it} \cdot \frac{\theta^M_{it}}{\theta^M_{jlt}} \cdot \mu_{it}^{-1}\right)^{-1}$, similar to Edmond, Midrigan and Xu (2019).

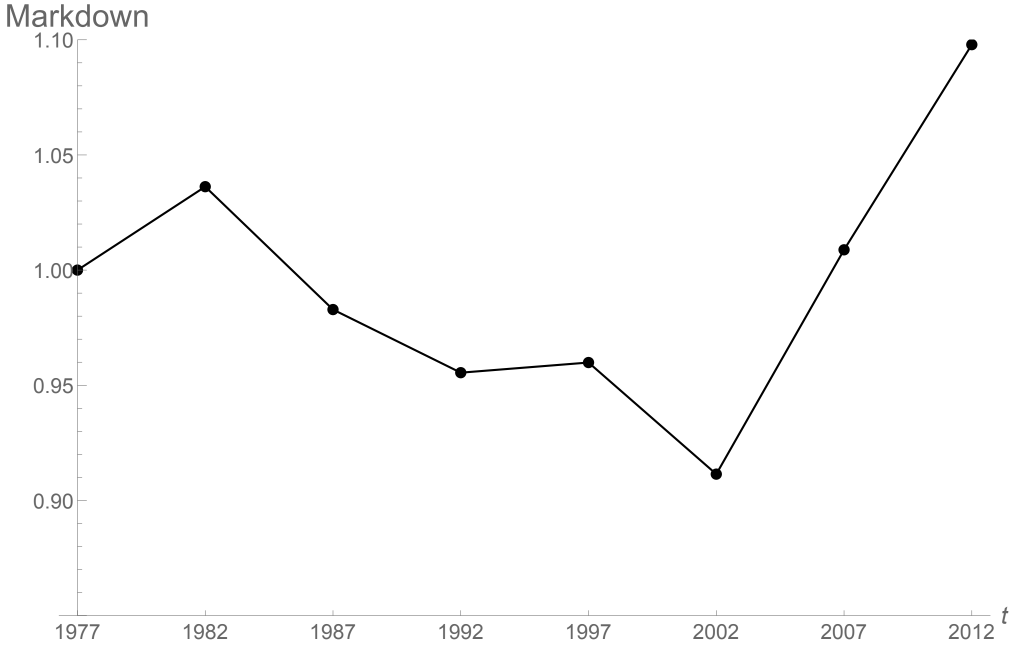
Using an analogous argument, we derive that the product of the aggregate markdown and markup is a weighted harmonic average of the product of firm-level markdowns and markups. We obtain:

$$\mathcal{V}_{jlt} \cdot \mathcal{M}_{jlt} = \left(\sum_{i \in F_t(j,l)} s_{it} \cdot \frac{\theta^L_{it}}{\theta^L_{jlt}} \cdot (\nu_{it}\mu_{it})^{-1}\right)^{-1}$$

Given that we have an expression for the aggregate markup, the expression for the aggregate markdown follows automatically. If output elasticities do not vary across firms within a given labor market, i.e. firms have Cobb-Douglas production technologies, then aggregation follows by taking sales-weighted harmonic averages. Whenever production technologies are not Cobb-Douglas, we need only apply correction terms that deal with heterogeneity in output elasticities — as can be seen from equation (11) in proposition 2.

We then aggregate across labor markets through employment weights. Finally, with some abuse of notation,

Figure 4: Time evolution of the aggregate markdown across U.S. manufacturing plants from 1977 to 2012.



Markdowns are constructed under the assumption of translog production and aggregated according to expression (11) and (13). The aggregate markdown is normalized relative to its initial value in 1977. Source: authors' own calculations from quinquennial CM data from 1977–2012.

we define the aggregate markdown as follows:

$$\mathcal{V}_t = \sum_{j \in J} \sum_{l \in L} \omega_{jlt} \mathcal{V}_{jlt} \tag{13}$$

where $\mathcal{V}_{jlt}$ is as in equation (11). Following the literature on markups (e.g., De Loecker and Eeckhout, 2018) and concentration (Autor et al., 2017; Rossi-Hansberg, Sarte and Trachter, 2018), we proceed by constructing markdowns at the firm, rather than plant, level using the CM.[20]

The resulting time trend of aggregate markdowns, as measured by $\mathcal{V}_t$, is depicted in figure 4. As opposed to previously documented trends on markups (e.g., De Loecker and Eeckhout, 2018), the aggregate markdown $\mathcal{V}_t$ is not monotonic. Instead, $\mathcal{V}_t$ is decreasing until the early 2000s, after which it begins to sharply increase. This time series pattern does not seem to support the notion that increasing labor market power by firms is the primary cause of the decline in the labor share, which began well before the early 2000s. The stark increase in the aggregate markdown since this time is potentially interesting, as several secular trends related to the decline in U.S. business dynamism have accelerated over the same horizon (see Decker et al., 2016).

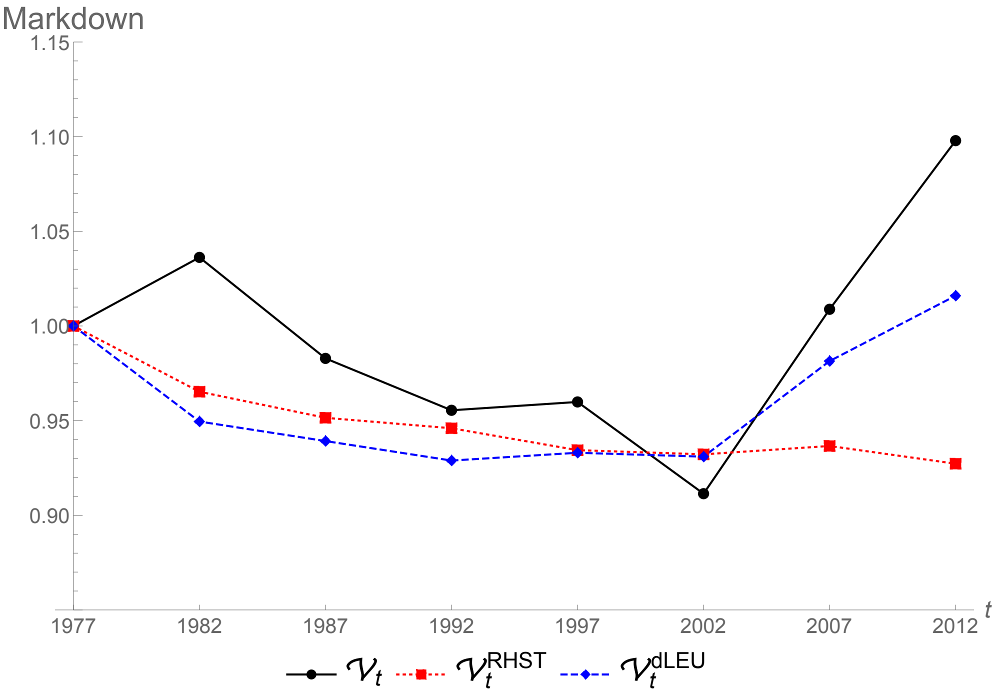
[20] By construction, the aggregate markdown is an employment-weighted average of markdowns at the market level. The latter is constructed using equations (11) and (12). However, it is difficult to construct these objects with the previously used ASM sample since our definition of a local labor market is rather narrow. Recall that the ASM is a representative sample and does not contain the universe of manufacturing plants. Hence, the number of observations that is used to construct $\mathcal{V}_{jlt}$ and $\mathcal{M}_{jlt}$ might be rather small for some labor markets (j, l) and induce biases in these objects from measurement error. To resolve this issue, we instead utilize the CM, which contains the universe of manufacturing plants but only at a quinquennial frequency.

Contrasting the time series for the aggregate markdown in equation (13) with two commonly used alternatives highlights the importance of using a local measure of aggregate markdowns that is also micro-founded. The first alternative we consider is the labor market equivalent of the aggregate markup measure used in De Loecker, Eeckhout and Unger (2018):

$$\begin{aligned}\mathcal{V}_t^{\text{dLEU}} &= \sum_{p\in P_t} \omega_{pt}\nu_{pt} \\ &= \sum_{f\in F_t} \omega_{ft}\left[\sum_{p\in P_t(f)} s_{pft}\nu_{pt}\right] \\ &\equiv \sum_{f\in F_t} \omega_{ft}\nu_{ft} \qquad (14)\end{aligned}$$

where P_t denotes the set of active plants in year t and s_{pft} the employment share of plant p in firm f. By construction, $\mathcal{V}_t^{\text{dLEU}}$ is an employment-weighted average of plant-level markdowns. This is identical to a firm-level average whenever firm-level markdowns are calculated as employment-weighted averages across a firm's plants.

Figure 5: The micro-founded aggregate markdown measure $\mathcal{V}_t$ (solid black) is decreasing between 1977 and 2002, and increasing afterwards. The employment-weighted aggregate markdown à la De Loecker, Eeckhout and Unger (2018) (dashed blue) shows a similar pattern, while a local aggregate inspired by local concentration in Rossi-Hansberg, Sarte and Trachter (2018) (dotted red) is monotonically decreasing.



Markdowns are constructed under the assumption of translog production and aggregated according to expressions equation (13), equation (14), and equation (15), respectively. All measures are normalized relative to their initial value in 1977. Source: authors' own calculations from quinquennial CM data from 1977–2012.

A second option is a measure for the aggregate markdown that mirrors the aggregate measure for local employment concentration, as in Rossi-Hansberg, Sarte and Trachter (2018). This approach still aggregates micro-level markdowns through employment weights, similarly to equation (14), but does so in two stages. First, micro-level markdowns are aggregated through their respective employment shares *within* each market, then markets are aggregated through employment weights to construct an aggregate measure. This leads us to:

$$\mathcal{V}_t^{\text{RHST}} = \sum_{j \in J} \sum_{l \in L} \omega_{jlt} M_{jlt} \text{ with } M_{jlt} = \sum_{f \in F_t(j,l)} \omega_{fjlt} \nu_{fjlt} \tag{15}$$

Figure 5 illustrates that, while our preferred measure $\mathcal{V}_t$ is decreasing until the early 2000s and sharply increasing afterwards, the alternatives display a different time evolution. While $\mathcal{V}_t^{\text{dLEU}}$ follows our measure in a qualitative sense, $\mathcal{V}_t^{\text{RHST}}$ is monotonically decreasing over the whole period.[21]

4.2 Comparing measures of market power with concentration

A number of recent studies have used measures of concentration as proxies for market power and inferred the prevalence and time evolution of market power from time trends in the concentration of either output or input markets. In this section, we discuss whether concentration is an accurate proxy for market power — at least within manufacturing — by comparing its cross-sectional and time-series properties with our estimated markdowns.

The Herfindahl-Hirschman index (HHI) is a canonical way to summarize the level of concentration in output markets — most notably, it is used in Autor et al. (2017) and Rossi-Hansberg, Sarte and Trachter (2018) and has been increasingly popular in studies of labor markets as well (see, for example, Azar, Marinescu and Steinbaum, 2017; Azar et al., 2018; Benmelech, Bergman and Kim, 2018; Rinz, 2018). However, it is ex-ante not clear whether concentration and market power should be correlated positively with each other in the first place. It may seem intuitive that large employers are able to exert the most labor market power, but it cannot be ruled out ex-ante that market power and concentration are instead negatively correlated.

As Syverson (2019) points out for output markets, this negative correlation is not "just a theoretical curiosity". It arises naturally in the framework of Melitz and Ottaviano (2008) and is at times observed in the data (as documented in Syverson, 2004a; Syverson, 2004b; and Goldmanis et al., 2010). Syverson (2019) concludes that caution should be exercised when using concentration to measure market power. The IO literature is most critical of using concentration measures as proxies for market power and has abandoned this practice

[21] Though the qualitative increase in $\mathcal{V}_t^{\text{dLEU}}$ holds for both translog (blue) and Cobb-Douglas (red) specifications, the differences between translog and Cobb-Douglas are quantitatively quite stark (as depicted in figure 7 in the appendix). These differences underline that Cobb-Douglas specifications can be quite restrictive. By construction, the Cobb-Douglas specification assumes that output elasticities are constant and, hence, ignores any time variation in a plant's output elasticities. Conversely, a translog specification allows precisely for this. Our results favor the translog specification as they indicate that this time variation is quantitatively important. This point is made even starker when focusing on the measure $\mathcal{V}_t^{\text{RHST}}$. If this measure is constructed under Cobb-Douglas markdowns, it is strongly increasing over time instead. Figure 8 in the appendix displays secular trends for aggregate markdowns when assuming Cobb-Douglas technologies instead.

after a short surge of the "structure-conduct-performance" literature. Despite this criticism, concentration indexes have never been explicitly compared to "ideal" measures of market power; at least not at a scale as wide as the whole manufacturing sector. This is precisely what we aim to do in this section.

For our comparison between aggregate markdowns and measures of concentration, we adopt the HHI as our main measure of market-level concentration and define it in a standard fashion:

$$\mathrm{HHI}_{mt} = \sum_{f \in F_t(m)} \left(\frac{x_{ft}}{X_{F(m)t}} \right)^2 \text{ s.t. } X_{F(m)t} = \sum_{f' \in F_t(m)} x_{f't} \tag{16}$$

where m denotes a market, $F_t(m)$ the set of firms in market m during a year t, and x is a measure of size (often employment or sales). We will mainly focus on labor markets and, hence, set $m = (j, \ell)$ to remain consistent with our previous analyses.

By construction, the HHI as defined in (16), ranges from $1/|F_t(m)|$ to 1 — where a value of 1 indicates maximum concentration, i.e. the presence of only one active seller/employer in a specific market-year. If firms were equally-sized, the inverse of the HHI would be equal to the number of employers $|F_t(m)|$ in a market m.

The literature has focused on two ways of combining market-level measures of concentration in order to construct a measure of *aggregate* concentration. Under the first approach, HHIs are constructed at the industry level (so that a market m is an industry) and are then aggregated through employment or sales weights. Following Autor et al. (2017), we refer to these as measures of national concentration.[22]

In contrast to the "national" approach, Rossi-Hansberg, Sarte and Trachter (2018) have argued that product market competition is better captured at the local level. Therefore, product (or labor) markets are defined through sector-location cells instead. Formally, an aggregate measure of local concentration is defined as:

$$\begin{aligned} \mathrm{LOCAL}_t &= \sum_{j \in J} \sum_{l \in L} \omega_{jlt} \mathrm{HHI}_{jlt} \qquad (17) \\ &= \sum_{j \in J} \sum_{l \in L} \omega_{jlt} \left[\sum_{f \in F_t(j,l)} \left(\frac{x_{flt}}{X_{F(j,l)t}} \right)^2 \right] \text{ s.t. } X_{F(j,l)t} = \sum_{f' \in F_t(j,l)} x_{f'lt} \end{aligned}$$

[22] More precisely, national concentration is constructed as:

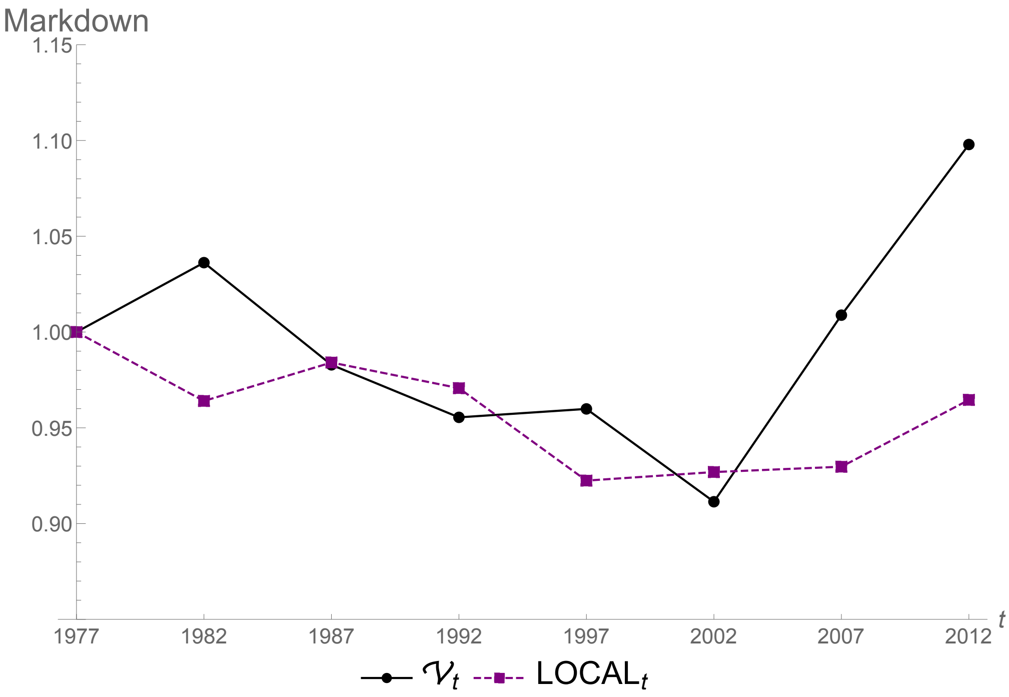
$$\mathrm{NATIONAL}_t = \sum_{j \in J} \omega_{jt} \mathrm{HHI}_{jt} = \sum_{j \in J} \omega_{jt} \left[\sum_{f \in F_t(j)} \left(\frac{x_{ft}}{X_{F(j)t}} \right)^2 \right] \text{ s.t. } X_{F(j)t} = \sum_{f' \in F_t(j)} x_{f't}$$

Autor et al. (2017) also provide aggregate national concentration measures based on the industry-level prominence of the top 4 or top 20 firms (also known as CR4 and CR20, respectively). Our conclusions remain unchanged if we use these CR measures instead.

Outcomes and weights can be defined through a variety of variables. In this paper, we implement equation (17) with data on employment as our preferred measure of aggregate concentration, following our reasoning on the local nature of labor markets as in section 4.1. Furthermore, we aim to compare the aggregate local concentration measure in equation (17) with several measures of the aggregate markdown.[23]

We start assessing the validity of HHIs as proxies for market power by calculating the cross-sectional correlations between employment concentration as in equation (17) and our preferred measure of local markdowns (as in equation (11)). These correlations indicate that more concentrated markets do not display higher markdowns on average, suggesting that the cross-sectional variation in local employment concentration is not necessarily reflecting variation in employer market power as measured by markdowns.

Figure 6: Comparing local markdowns with local employment concentration: the time evolution in local employment concentration is consistent with broad trends in labor market power in U.S. manufacturing but is somewhat imprecise with respect to both timing and magnitude.



The solid line depicts the time series for local markdowns as in equation (11), whereas the purple, dashed line shows local employment concentration as in equation (17). All measures are normalized relative to their initial value in 1977. Source: authors' own calculations from quinquennial CM data from 1977 – 2012.

The comparison between aggregate local concentration and aggregate markdowns over time yields slightly

[23] Rossi-Hansberg, Sarte and Trachter (2018) adopt a local concentration measure akin to (17) through sales HHI and employment weights in the National Establishment Time Series (NETS) data. Similarly, Rinz (2018) has constructed local employment concentration measures with HHIs and weights defined through employment using Census data. We have used a variety of data sources on firm-level employment and vacancies to explore different approaches to computing aggregate labor market concentration and how they relate to each other. While the main body of the paper presents results only for (stock) employment concentration, results based on vacancies, job creation flows, or payroll depict a similar picture when compared to employment count. The results for vacancies, based on data from Burning Glass Technologies (BGT), can be found in Appendix F. The remaining results are available upon request.

more reassuring results. Figure 6 displays the time trends for our preferred aggregate markdown measure $\mathcal{V}_t$ and local employment concentration LOCAL_t. Although both series generally decline over time until the early 2000s, the aggregate markdown experiences a sharp trend inversion at this time that is not reflected in a similar change in employment concentration until several years later, during the Great Recession. Furthermore, the increase in local concentration we observe since 2007 is much smaller than the increase documented for markdowns since the early 2000s. In conclusion, the evolution in local employment concentration is roughly consistent with broad trends in labor market power in U.S. manufacturing, but it is somewhat imprecise with respect to both timing and magnitude, indicating the two measures are imperfect substitutes.[24]

5 Conclusion

This paper provides a characterization of employer market power in the U.S. manufacturing sector, both in the cross-section and over time. We start by estimating markdowns — the wedge between marginal revenue products of labor and wages — at the plant-year level using the "production approach". We find that labor markets in U.S. manufacturing are far from perfectly competitive: the average plant operates in a monopsonistic environment, as it charges a markdown of 1.53. In other words, a worker employed at the average manufacturing plant earns 65 cents of each dollar generated on the margin. We also document that there is a substantial amount of dispersion in markdowns. In our whole sample, the interquartile range of markdowns is equal to 61.8 percent. More importantly though, most of the markdown variation is observed within detailed industries, with an average within-industry interquartile range (standard deviation) of 61.6 (60.4) percent. Furthermore, we find that plant size, as well as sectoral and geographical scope, are also associated with greater markdowns, although we find less of a role for revenue productivity or being in a high-tech industry.

We also investigate long-term trends in employer market power, via a measure of aggregate markdowns consistent with aggregate wedges and that accounts for local labor markets. Importantly, aggregation occurs through sales-weighted harmonic averages that adjust for production heterogeneity across firms. We find aggregate markdowns decreased between the late 1970s and early 2000s but increased sharply afterward. This non-monotonic pattern is inconsistent with the view that the decline in the U.S. labor share (or wage stagnation) were induced by changes in labor market power. Furthermore, we show that popular measures of employment concentration do not line up well with the aggregate markdown, suggesting caution should be exercised when employment concentration is used as a proxy for employer market power.

While we believe our approach makes significant strides in the estimation and trend measurement of markdowns, we have only scratched the surface in understanding how and why markdowns vary. For example, we do not yet know whether the cross-industry variation in markdowns can be rationalized by industry-level observables such as unionization rates, non-compete agreements, labor regulations (e.g., right-to-work laws),

[24] Appendix E reports cross-sectional tables and time trends for the national concentration measure as well. Our conclusion that concentration is a poor proxy for employer market power is unchanged when using national measures.

or import competition. Such empirical exercises could help us further understand the determinants — and welfare implications — of employer market power.

We also acknowledge our approach is not without shortcomings. While it is compatible with a broad array of monopsony frameworks, it rules out any model of monopsony in which firms' market power does not originate from an upward-sloping labor supply curve. Most notably, our results cannot be interpreted through the lens of models in the family of Diamond (1982) and Mortensen and Pissarides (1994). However, Dobbelaere and Mairesse (2013) show that wedges between output elasticities and revenue shares can also be used to identify *firm*-level parameters of a static Nash bargaining problem in which risk-neutral workers and firms negotiate over wages and the level of employment. These estimated parameters can be informative for characterizing employer market power in random search settings with perfectly elastic labor supply curves. We leave investigation of these themes for future research.

References

Abowd, J.M., F. Kramarz, and D.N. Margolis, "High Wage Workers and High Wage Firms," *Econometrica*, 1999, *67* (2), 251 – 333.

Ackerberg, D., K. Caves, and G. Frazer, "Identification Properties of Recent Production Function Estimators," *Econometrica*, 2015, *83* (6), 2411 – 2451.

Amior, M. and A. Manning, "The Persistence of Local Joblessness," *American Economic Review*, July 2018, *108* (7), 1942–1970.

Ashenfelter, O.C., H. Farber, and M.R. Ransom, "Labor Market Monopsony," in A. Manning, ed., *Journal of Labor Economics*, University of Chicago Press, 2011, pp. 203 – 2010.

Atalay, E., "Materials Prices and Productivity," *Journal of the European Economic Association*, 2014, *12* (3), 575 – 611.

Atkeson, A. and A. Burstein, "Pricing-to-Market, Trade Costs, and International Relative Prices," *American Economic Review*, 2008, *98* (5), 1998 – 2031.

Autor, D., D. Dorn, L.F. Katz, C. Patterson, and J. Van Reenen, "The Fall of the Labor Share and the Rise of Superstar Firms," NBER Working Paper, 2017, (23396).

Azar, J., O. Marinescu, and M.I. Steinbaum, "Labor Market Concentration," NBER Working Paper, 2017, (24147).

_ , _ , _ , and B. Taska, "Concentration in U.S. Labor Markets: Evidence from Online Vacancy Data," NBER Working Paper, 2018, (24395).

Baqaee, D. and E. Farhi, "Productivity and Misallocation in General Equilibrium," Working Paper, 2018.

Basu, S., "Intermediate Goods and Business Cycles: Implications for Productivity and Welfare," *American Economic Review*, 1995, *85* (3), 512 – 531.

_ and J.G. Fernald, "Returns to Scale in U.S. Production: Estimates and Implications," *Journal of Political Economy*, 1997, *105* (2), 249 – 283.

Benmelech, E., N. Bergman, and H. Kim, "Strong Employers and Weak Employees: How Does Employer Concentration Affect Wages?," NBER Working Paper, 2018, (24307).

Berger, D., K. Herkenhoff, and S. Mongey, "Labor Market Power," NBER Working Paper, 2019, (25719).

Bhaskar, V. and T. To, "Minimum Wages for Ronald McDonald Monopsonies: A Theory of Monopsonistic Competition," *Economic Journal*, 1999, *109*, 190 – 203.

Boal, W.M. and M.R. Ransom, "Monopsony in the Labor Market," *Journal of Economic Literature*, 1997, *35*, 86 – 112.

Bontemps, C., J.-M. Robin, and G.J. Van den Berg, “An Empirical Equilibrium Job Search Model with Search on the Job and Heterogeneous Workers and Firms,” *International Economic Review*, 2001, *40* (4), 1039 – 1074.

Brooks, W., J. Kaboski, Y. A. Li, and W. Qian, “Exploitation of Labor? Classical Monopsony Power and Labor’s Share,” Working Paper, 2019.

Burdett, K. and D.T. Mortensen, “Wage Differentials, Employer Size, and Unemployment,” *International Economic Review*, 1998, *39* (2), 257 – 273.

Caldwell, S. and E. Oehlsen, “Monopsony and the Gender Wage Gap: Experimental Evidence from the Gig Economy,” Working Paper, 2018.

Card, D., A.R. Cardoso, J. Heining, and P. Kline, “Firms and Labor Market Inequality: Evidence and Some Theory,” *Journal of Labor Economics*, 2018, *36* (1), 13 – 70.

_ , F. Devicienti, and A. Maida, “Rent-Sharing, Holdup, and Wages: Evidence from Matched Panel Data,” *Review of Economic Studies*, 2014, *81* (1), 84–111.

Chan, M., K. Kroft, and I. Mourifie, “An Empirical Framework for Matching with Imperfect Competition,” Working Paper, 2019.

Christofides, L.N. and A.J. Oswald, “Real Wage Determination and Rent-Sharing in Collective Bargaining Agreements,” *Quarterly Journal of Economics*, 1992, *107* (3), 985–1002.

Cole, H. and L. Ohanian, “The U.S. and U.K. Great Depressions through the Lens of Neoclassical Growth Theory,” *American Economic Review*, 2002, *92* (2), 28–32.

Cooper, R., J. Haltiwanger, and J.L. Willis, “Search frictions: Matching aggregate and establishment observations,” *Journal of Monetary Economics*, 2007, *54*, 56 – 78.

Davis, S. J., C. Grim, J. Haltiwanger, and M. Streitwieser, “Electricity Unit Value Prices and Purchase Quantities: U.S. Manufacturing Plants, 1963 - 2000,” *Review of Economics and Statistics*, 2013, *95* (4), 1150 – 1165.

Davis, S.J., R.J. Faberman, and J. Haltiwanger, “Labor market flows in the cross-section and over time,” *Journal of Monetary Economics*, 2012, *59* (1), 1 – 18.

Decker, R.A., J. Haltiwanger, R.S. Jarmin, and J. Miranda, “Declining business dynamism: what we know and the way forward,” *American Economic Review: Papers and Proceedings*, 2016, *106* (5), 203 – 2007.

Diamond, P.A., “Wage Determination and Efficiency in Search Equilibrium,” *Review of Economic Studies*, 1982, *49*, 217 – 227.

Dobbelaere, S. and J. Mairesse, “Panel data estimates of the production function and product and labor market imperfections,” *Journal of Applied Econometrics*, January 2013, *28* (1), 1–46.

Dube, A., J. Jacobs, S. Naidu, and S. Siddarth, "Monopsony in Online Labor Markets," NBER Working Paper, 2018, (24416).

Edmond, C., V. Midrigan, and D.Y. Xu, "How Costly Are Markups?," Working Paper, 2019.

Eggertsson, G., J.A. Robbins, and E.G. Wold, "Kaldor's and Piketty's Facts: The Rise of Monopoly Power in the United States," NBER Working Paper, 2018, (24287).

Elsby, M., B. Hobijn, and A. Sahin, "The Decline of the U.S. Labor Share," *Brookings papers on Economic Activity*, 2013, pp. 1 – 52.

Falch, T., "The Elasticity of Labor Supply at the Establishment Level," *Journal of Labor Economics*, 2010, *28* (2).

Flynn, Z., A. Gandhi, and J. Traina, "Identifying Market Power in Production Data," Working Paper, 2019.

Foster, L., J. Haltiwanger, and C. Syverson, "Selection on Productivity or Profitability?," *American Economic Review*, 2008, *98* (1), 394 – 425.

Gali, J., M. Gertler, and D.J. Lopez-Salido, "Markups, gaps, and the welfare costs of business fluctuations," *Review of Economics and Statistics*, 2007, *89* (1), 44–59.

Gandhi, A., S. Navarro, and D. Rivers, "On the Identification of Gross Output Production Functions," Working Paper, 2017.

Goldmanis, M., A. Hortacsu, C. Syverson, and O. Emre, "E-commerce and the Market Structure of Retail Industries," *Economic Journal*, 2010, *120* (545), 651 – 682.

Goolsbee, A. and C. Syverson, "Monopsony Power in Higher Education: A Tale of Two Tracks," Becker Friedman Institute for Economics Working Paper, 2019, (2019-95).

Hall, R. E., "The Relation between Price and Marginal Cost in U.S. Industry," *Journal of Political Economy*, 1988, *96* (5), 921 – 947.

Hall, R.E., "Measuring Factor Adjustment Costs," *Quarterly Journal of Economics*, 2004, *119*, 899 – 927.

Haltiwanger, J., R.S. Jarmin, and J. Miranda, "Who Creates Jobs? Small versus Large versus Young," *Review of Economics and Statistics*, 2013, *95* (2), 347 – 361.

Hsieh, C.-T. and P.J. Klenow, "Misallocation and Manufacturing TFP in China and India," *Quarterly Journal of Economics*, 2009, *74* (4), 1403 – 1448.

Huckfeldt, C., "Understanding the Scarring Effects of Recessions," Working Paper, 2017.

Itskhoki, O. and B. Moll, "Optimal Development Policies with Financial Frictions," *Econometrica*, 2019, *87* (1), 139 – 173.

Jarosch, G., J.S. Nimczik, and I. Sorkin, "Granular Search, Market Structure, and Wages," Working Paper, 2019, (26239).

Karabarbounis, L., "The labor wedge: MRS vs. MPN," *Review of Economic Dynamics*, 2014, *17*, 206–223.

_ and B. Neiman, "The Global Decline of the Labor Share," *Quarterly Journal of Economics*, 2013, *129* (1), 61 – 103.

Kehrig, M., "The Cyclical Nature of the Productivity Distribution," Working Paper, 2015.

Kennan, J. and J.R. Walker, "The Effect of Expected Income on Individual Migration Decisions," *Econometrica*, January 2011, *79* (1), 211–251.

Kim, R., "Price-Cost Markup Cyclicality: New Evidence and Implications," Working Paper, 2017. New York (NY): Columbia University.

Kimball, M.S., "The Quantitative Analytics of the Basic Neomonetarist Model," *Journal of Money, Credit and Banking*, 1995, *27* (4), 1241 – 1277.

Lamadon, T., M. Mogstad, and B. Setzler, "Imperfect Competition, Compensating Differentials and Rent Sharing in the U.S. Labor Market," Working Paper 25954, National Bureau of Economic Research June 2019.

Lazear, E.P. and J.R. Spletzer, "Hiring, Churn, and the Business Cycle," *American Economic Review*, 2012, *102* (3), 575 – 579.

Levinsohn, J.A. and A. Petrin, "Estimating Production Functions Using Inputs to Control for Unobservables," *Review of Economic Studies*, 2003, *70* (2), 317 – 340.

Lipsius, B., "Labor Market Concentration Does Not Explain the Falling Labor Share," Working Paper, 2018.

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

_ and F. Warzynski, "Markups and Firm-Level Export Status," *American Economic Review*, 2012, *102* (6), 2437 – 2471.

_ and J. Eeckhout, "Global Market Power," Working Paper, 2018.

_ , _ , and G. Unger, "The Rise of Market Power and the Macroeconomic Implications," Working Paper, 2018.

Macaluso, C., "Skill Remoteness and Post-Layoff Labor Market Outcomes," 2017, (Thesis), 100.

Manning, A., *Monopsony in Motion*, Princeton University Press, 2003.

_ , "Imperfect Competition in the Labor Market," in O. Ashenfelter and D. Card, eds., *Handbook of Labor Economics*, Vol. 4, North-Holland: Elsevier, 2011, chapter 11, pp. 973 – 1041.

_ **and B. Petrongolo**, “How Local Are Labor Markets? Evidence from a Spatial Job Search Model,” *American Economic Review*, 2017, *107* (10), 2877 – 2907.

Marinescu, O. and R. Rathelot, “Mismatch Unemployment and the Geography of Job Search,” *American Economic Journal: Macroeconomics*, 2018, *10* (3).

Matsudaira, J.D., “Monopsony in the Low-Wage Labor Market? Evidence from Minimum Nurse Staffing Regulations,” *Review of Economics and Statistics*, 2014, *96* (1).

Melitz, M.J. and G.I.P. Ottaviano, “Market Size, Trade, and Productivity,” *Review of Economic Studies*, 2008, *75* (1), 295 – 316.

Morlacco, M., “Market Power in Input Markets: Theory and Evidence from French Manufacturing,” Working Paper, 2019.

Mortensen, D., “How Monopsonistic is the (Danish) Labor Market?,” in P. Aghion, R. Frydman, J. Stiglitz, and M. Woodford, eds., *Knowledge, information and expectations in modern macroeconomics*, Princeton, NJ: Princeton University Press, 2003.

Mortensen, D.T. and C.A. Pissarides, “Job Creation and Job Destruction in the Theory of Unemployment,” *Review of Economic Studies*, 1994, *61*, 397 – 415.

Ransom, M.R. and D.P. Sims, “Estimating the Firm’s Labor Supply Curve in a “New Monopsony” Framework: Schoolteachers in Missouri,” *Journal of Labor Economics*, 2010, *28* (2).

Reenen, J. Van, “The Creation and Capture of Rents: Wages and Innovation in a Panel of U.K. Companies,” *Quarterly Journal of Economics*, 1996, *111* (1), 195–226.

Rinz, K., “Labor market Concentration, Earnings Inequality, and Earnings Mobility,” *CARRA Working Paper Series 2018-10*, 2018. Washington, DC: Center for Administrative Records Research and Applications.

Robinson, J., *The Economics of Imperfect Competition*, Palgrave Macmillan, 1933.

Rossi-Hansberg, E., P.-D. Sarte, and N. Trachter, “Diverging Trends in National and Local Concentration,” NBER Working Paper, 2018, (25066).

Salop, S., “Monopolistic Competition with Outside Goods,” *Bell Journal of Economics*, 1979, *10*, 141 – 156.

Sokolova, A. and T. Sorensen, “Monopsony in Labor Markets: A Meta-Analysis,” IZA Discussion Paper, 2018, (11966).

Staiger, D.O., J. Spetz, and C.S. Phibbs, “Is There Monopsony in the Labor Market? Evidence from a Natural Experiment,” *Journal of Labor Economics*, 2010, *28* (2).

Syverson, C., “Market Structure and Productivity: A Concrete Example,” *Journal of Political Economy*, 2004, *112* (6), 1181 – 1222.

_ , "Product Substitutability and Productivity Dispersion," *Review of Economics and Statistics*, 2004, *86* (2), 534 – 550.

_ , "Macroeconomics and Market Power: Facts, Potential Explanations and Open Questions," *Economic Studies at Brookings*, 2019.

Traina, J., "Is Aggregate Market Power Increasing? Production Trends Using Financial Statements," *Stigler Center New Working Paper Series*, 2018, (17).

Webber, D., "Firm market power and the earnings distribution," *Labour Economics*, 2015, *35* (C), 123 – 134.

A Estimation procedure for markdowns

A.1 Derivations

In this appendix, we formalize our arguments in the main text. In particular, we show that retrieving output elasticities and revenue shares are sufficient in order to estimate markdowns. To see this, we start with the cost minimization problem of a firm. In general, we have:

$$\min_{\mathbf{X}_{it}\in\mathbb{R}^K_+} \sum_{k=1}^{K} v^k_{it}(x^k_{it})x^k_{it} + \Phi^k\left(x^k_i, x^k_{it-1}\right) \quad \text{s.t.} \quad F(\mathbf{X}_{it};\omega_{it}) \geq Q_{it} \tag{18}$$

where $\mathbf{X}_{it} = \left(x^1_{it}, \ldots, x^K_{it}\right)'$ is the firm's vector of $K \geq 1$ production inputs with prices $\{v^k_{it}\}^K_{k=1}$. Furthermore, ω_{it} denotes a firm i's productivity level at time t. Adjustment costs for some input k are captured by the function $\Phi^k(\cdot,\cdot)$.

To derive markdowns, we start with the insight by Hall (1988) that the wedge between a flexible input's output elasticity and its revenue share must reflect a firm's output market power (or its mark*up*; defined as its output price over marginal cost of production). This result hinges upon the following assumptions.

A1. A firm i chooses its inputs according to COST MINIMIZATION problem (18).

There exists at least one input k' that features:

A2. FULL FLEXIBILITY. $\Phi^{k'}(\cdot,\cdot) = 0$.

A3. NO MONOPSONY POWER. $v^{k'}_{it}(x^{k'}_{it}) = v^{k'}_{it}$.

A4. SMOOTHNESS. $F(\cdot;\omega_{it})$ is twice continuously differentiable in $x^{k'}_{it}$ for any $\omega \in \mathbb{R}_+$.

A5. REGULARITY. $F(\cdot;\omega_{it})$ satisfies $\lim\limits_{x^{k'}_{it}\to 0} \dfrac{\partial F(\mathbf{X}_{it};\omega_{it})}{\partial x^{k'}_{it}} = +\infty$ and $\lim\limits_{x^{k'}_{it}\to +\infty} \dfrac{\partial F(\mathbf{X}_{it};\omega_{it})}{\partial x^{k'}_{it}} = 0$.

LEMMA 1. Let assumptions **A1** – **A5** hold, then a firm i's markup satisfies:

$$\begin{aligned}\mu_{it} &= \frac{\partial F(\mathbf{X}_{it};\omega_{it})}{\partial x^{k'}_{it}} \frac{x^{k'}_{it}}{Q_{it}} \cdot \left(\frac{v^{k'}_{it} x^{k'}_{it}}{p_{it}Q_{it}}\right)^{-1} \\ &\equiv \frac{\theta^{k'}_{it}}{\alpha^{k'}_{it}}\end{aligned} \tag{19}$$

for any flexible input k'.

Proof. Under the stated assumptions, the first order condition for any flexible input k', associated with cost minimization problem (18), satisfies:

$$v_{it}^{k'} = \lambda_{it} \frac{\partial F(\mathbf{X}_{it}; \omega_{it})}{\partial x_{it}^{k'}}$$

where λ_{it} is the Lagrangian multiplier associated with cost minimization problem (18). This shadow value of total variable costs is also known as firm i's marginal cost of production. The above equality can easily be manipulated to:

$$\frac{v_{it}^{k'} x_{it}^{k'}}{p_{it} Q_{it}} = \frac{\lambda_{it}}{p_{it}} \frac{\partial F(\mathbf{X}_{it}; \omega_{it})}{\partial x_{it}^{k'}} \frac{x_{it}^{k'}}{Q_{it}}$$

where p_{it} denotes a firm's price for its output good. Then, we get the expression for a firm i's markup $\mu_{it} = \frac{p_{it}}{\lambda_{it}}$ at time t:

$$\mu_{it} = \frac{\theta_{it}^{k'}}{\alpha_{it}^{k'}} \tag{20}$$

where $\theta_{it}^{k'} \equiv \frac{\partial F(\mathbf{X}_{it}; \omega_{it})}{\partial x_{it}^{k'}} \frac{x_{it}^{k'}}{Q_{it}}$ and $\alpha_{it}^{k'} \equiv \frac{v_{it}^{k'} x_{it}^{k'}}{p_{it} Q_{it}}$. Thus, a firm's markup is equal to the wedge between the output elasticity and the revenue share of some input k'. Note that the existence of *only one* flexible input k' is sufficient to establish this result. □

Any input k' that satisfies assumptions **A2**, **A3**, **A4** and **A5** simultaneously, will be referred to as a flexible input. Assumptions **A1**, **A4** and **A5** are weak, so the challenge becomes to find an input that satisfies assumptions **A2** and **A3** simultaneously. In plain words, this requires the existence of an input k' that is free of adjustment costs and firms are price-takers for this particular input.

PROPOSITION 3. Let assumptions **A1** – **A5** hold for some input k' that is not equal to labor. In addition, let assumptions **A4** and **A5** apply to labor ℓ. If a firm i faces a differentiable, finitely-elastic wage schedule $w(\ell)$ and labor is chosen statically, then its markdown satisfies:

$$\nu_{it} = \frac{\theta_{it}^{\ell}}{\alpha_{it}^{\ell}} \cdot \left(\frac{\theta_{it}^{k'}}{\alpha_{it}^{k'}} \right)^{-1} \tag{21}$$

Proof. Without loss of generality, consider the following *conditional* cost minimization problem:

$$\min_{\ell_{it} \geq 0} w_{it}(\ell_{it}) \ell_{it} \text{ s.t. } F(\ell_{it}, \mathbf{X}^*_{-\ell,it}; \omega_{it}) \geq Q_{it}$$

The associated optimality condition with Lagrangian multiplier λ_{it} can be characterized as:

$$\left[\frac{w'_{it}(\ell_{it})\ell_{it}}{w_{it}(\ell_{it})}+1\right]=\lambda_{it}\cdot\frac{\frac{\partial F(\ell_{it},\mathbf{X}^*_{-\ell,it};\omega_{it})}{\partial \ell_{it}}}{w_{it}(\ell_{it})}$$

which we can rearrange as:

$$\begin{aligned}\left[\frac{w'_{it}(\ell_{it})\ell_{it}}{w_{it}(\ell_{it})}+1\right]&\equiv\varepsilon_S^{-1}(\ell_{it})+1\\&=\frac{\lambda_{it}}{p_{it}}\cdot\frac{\partial F(\ell_{it},\mathbf{X}^*_{-\ell,it};\omega_{it})}{\partial \ell_{it}}\frac{\ell_{it}}{Q_{it}}\cdot\frac{p_{it}Q_{it}}{w_{it}(\ell_{it})\ell_{it}}\\&\equiv\mu_{it}^{-1}\cdot\frac{\theta^{\ell}_{it}}{\alpha^{\ell}_{it}}\end{aligned}\tag{22}$$

Given our insight on a firm’s markdown, we must have:

$$\frac{\theta^{\ell}_{it}}{\alpha^{\ell}_{it}}=\nu_{it}\cdot\mu_{it}\tag{23}$$

Then, the result follows immediately from lemma 1. Hence, we have:

$$\nu_{it}=\frac{\theta^{\ell}_{it}}{\alpha^{\ell}_{it}}\cdot\left(\frac{\theta^{k'}_{it}}{\alpha^{k'}_{it}}\right)^{-1}\tag{24}$$

which is what we wanted to show. □

Note that the result from the main text follows immediately from the above proposition whenever material inputs are assumed to be flexible. The revenue shares α^L_{it} and α^M_{it} can be directly constructed from the data. To obtain markdowns (and markups), it is sufficient to estimate output elasticities only. Therefore, we need to estimate production functions.

A.2 GMM-IV estimation procedure

In the following, we will provide more details on how we obtain output elasticities. To do so, we will follow the “proxy variable” literature on production function estimation (Levinsohn and Petrin, 2003;De Loecker and Warzynski, 2012; Ackerberg, Caves and Frazer, 2015). This literature imposes a set of additional assumptions to obtain firms’ production technologies. Before we list these assumptions formally, we need to introduce some notation.

Let the production function be given by:

$$Q_{it}=F(\mathbf{V}_{it},\mathbf{K}_{it};\omega_{it})$$

where we categorize inputs as flexible or non-flexible inputs, i.e. $\mathbf{X}'_{it} = (\mathbf{V}'_{it}, \mathbf{K}'_{it})$. In particular, we have:

$$\mathbf{V}_{it} = (V^1_{it}, \dots, X^V_{it})'$$
$$\mathbf{K}_{it} = (K^{V+1}_{it}, \dots, K^K_{it})'$$

where the first $V \geq 1$ inputs are flexible and the latter $K - V + 1$ inputs are not fully flexible. $\mathbf{K}_{it}$ is a state variable when choosing the inputs $\mathbf{V}_{it}$. Furthermore, ω_{it} denotes a plant's productivity. In particular, suppose that $X^1_{it} = M_{it}$ are material inputs. The following set of assumptions allows us to estimate output elasticities:

A6. Technology parameters are constant across time and common within an industry group.

A7. MONOTONICITY. The determinant $D^M \in \mathbb{R}^{V \times V}$ **globally** satisfies:

$$D^M = \left| \frac{\partial \mathbf{V}_{it}(\mathbf{K}_{it}, \omega_{it})}{\partial \omega_{it}} \quad \mathbf{H}^F_{2,V}(\mathbf{K}_{it}, \omega_{it}) \quad \dots \quad \mathbf{H}^F_{V,V}(\mathbf{K}_{it}, \omega_{it}) \right| > 0$$

where $\frac{\partial \mathbf{V}_{it}(\mathbf{K}_{it}, \omega_{it})}{\partial \omega_{it}} = \left(\frac{\partial X^1_{it}(\mathbf{K}_{it}, \omega_{it})}{\partial \omega_{it}}, \dots, \frac{\partial X^V_{it}(\mathbf{K}_{it}, \omega_{it})}{\partial \omega_{it}} \right)'$ and
$\mathbf{H}^F_{r,V}(\mathbf{K}_{it}, \omega_{it}) = \left(\frac{\partial F(\mathbf{V}_{it}, \mathbf{K}_{it}, \omega_{it})}{\partial X^r_{it} \partial X^1_{it}}, \dots, \frac{\partial F(\mathbf{V}_{it}, \mathbf{K}_{it}, \omega_{it})}{\partial X^r_{it} \partial X^V_{it}} \right)'$ is the r^{th} column of the Hessian matrix for $F(\cdot, \mathbf{K}_{it}; \omega_{it})$ evaluated at $\mathbf{V}_{it} \in \mathbb{R}^V_+$.

A8. Productivity ω is one-dimensional and HICKS-NEUTRAL.

A9. MARKOV PROPERTY. Plant-level productivity ω_{it} follows a stochastic process of the form:

$$\omega_{it} = g_t(\omega_{it-1}) + \xi_{it}$$

where the technology shock satisfies $\mathbb{E}_{t-1}(\xi_{it} | \omega_{it-1}) = 0$.

Assumption **A6** is without much loss of generality as we can allow for technology parameters that are time-varying and/or common within an industry-geographic cell group instead. Assumption **A7** is a sufficient condition for material inputs to be invertible in productivity and can be seen as the generalized version of the main assumption in appendix A of Levinsohn and Petrin (2003).

While assumptions **A8** and **A9** are not fully general, the vast majority of frameworks in firm dynamics are nested by these two assumptions. This is because we allow for AR(1) processes for one-dimensional levels of productivity. With some abuse of notation, assumptions **A6** and **A8** imply:

$$F(\mathbf{V}_{it}, \mathbf{K}_{it}; \omega_{it}) = F(\mathbf{V}_{it}, \mathbf{K}_{it}; \boldsymbol{\beta})\text{exp}(\omega_{it})$$

where $\boldsymbol{\beta} \in \mathbb{R}^Z$ parameterizes the production function. Note that assumption **A6** neither implies that output elasticities are constant over time nor that they cannot differ across industries.

To account for measurement error, we assume that *observed* logged output satisfies $y_{it} = \ln(Q_{it}) + \epsilon_{it}$, i.e. measurement error enters production in a multiplicative fashion. Note that the error term ϵ_{it} is *not* observed by firms when they have to make their optimal input decisions. All in all, we have:

$$y_{it} = f(\mathbf{v}_{it}, \mathbf{k}_{it}; \boldsymbol{\beta}) + \omega_{it} + \epsilon_{it}$$

where $f(\mathbf{v}_{it}, \mathbf{k}_{it}; \boldsymbol{\beta}) = \ln\left(F(\mathbf{V}_{it}, \mathbf{K}_{it}; \boldsymbol{\beta})\right)$, and $\mathbf{v}_{it}$ and $\mathbf{k}_{it}$ denote componentwise natural log transformations of $\mathbf{V}_{it}$ and $\mathbf{K}_{it}$, respectively. Firm-level productivities ω_{it} are not observed by the econometrician, but are observable for firms themselves.

Unobservable productivity is the main cause of endogeneity concerns in our estimation procedure. To deal with this, we use the insight initiated by Levinsohn and Petrin (2003). Under assumption **A7**, material demand $\ln\big(X_{it}^1\big) = m_{it}$ can be used to proxy for productivity. Note that plants choose flexible inputs given the state $\mathbf{K}_{it}$, idiosyncratic productivity ω_{it} and some controls that can influence their decisions $\mathbf{c}_{it}$ (e.g., input prices):

$$m_{it} = m_t(\omega_{it}; \mathbf{k}_{it}, \mathbf{c}_{it})$$

where the vector $\mathbf{c}_{it}$ denotes any additional variables that can affect a plant's optimal demand for material inputs.[25] The above mapping for materials is invertible in productivity ω_{it} whenever assumption **A7** holds. Under this assumption, the material input demand function is monotonic in productivity ω_{it}.[26] Then, there exists some function $h_t(\cdot; \mathbf{k}_{it}, \mathbf{c}_{it})$ such that:

$$\omega_{it} = h_t(m_{it}, \mathbf{k}_{it}, \mathbf{c}_{it})$$

As a result, production y_{it} can be written in terms of observables only:

$$\begin{aligned} y_{it} &= f(\mathbf{v}_{it}, \mathbf{k}_{it}; \boldsymbol{\beta}) + h_t(m_{it}; \mathbf{k}_{it}, \mathbf{c}_{it}) \\ &= \phi_t(\mathbf{v}_{it}, \mathbf{k}_{it}, \mathbf{c}_{it}) + \epsilon_{it} \\ &= \varphi_{it} + \epsilon_{it} \end{aligned}$$

Estimating the production technology parameters $\boldsymbol{\beta}$ is done in a three stage fashion which is in a similar spirit to Ackerberg, Caves and Frazer (2015). To implement our estimation procedure, we set $\mathbf{v}_{it} = m_{it}$, $\mathbf{k}_{it} = (k_{it}, \ell_{it}, e_{it})'$ and $\mathbf{c}_{it} = (d_{i,1}, \ldots, d_{i,T})'$ where $d_{i,t}$ is a fixed effect for a specific year t. Even though

[25] In the empirical implementation, $\mathbf{c}_{it}$ only contains a set of year fixed effects. Industry fixed effects are not required whenever production technology parameters are estimated industry-by-industry. However, the used methodology is flexible enough to account for other observables.

[26] This follows from standard arguments for comparative statics under multiple inputs. We then apply Cramer's rule to arrive at the condition summarized in assumption **A7**. Levinsohn and Petrin (2003) show a similar result for $V = 2$ in their appendix A. In a nutshell, assumption **A7** imposes a set of regularity conditions on the cross-derivatives of the production function in $\mathbf{V}_{it}$ which are fairly mild.

we will mainly focus on translog production functions, we also occasionally report results for Cobb-Douglas specificiations.

Step 1. Non-parametric estimation of φ_{it} and ϵ_{it}.

First, we estimate φ_{it} and ϵ_{it} non-parametrically by a third degree polynomial in $\mathbf{x}_{it} = (k_{it}, \ell_{it}, m_{it}, e_{it})'$ with interaction terms. Let its fitted values and residuals be denoted by $\widehat{\varphi}_{it}$ and $\widehat{\epsilon}_{it}$ respectively. These residuals are then interpreted as measurement error in observed output.

Step 2. Construction of innovations ξ_{it} to productivity ω_{it}.

By assumption **A9**, idiosyncratic productivity ω_{it} is Markovian, thus its expected value is only a function of its lagged value. As a result, we have $\omega_{it} = g_t(\omega_{it-1}) + \xi_{it}$. Then, productivity is approximated in the data by:

$$\omega_{it}(\boldsymbol{\beta}) = \widehat{\varphi}_{it} - f(\mathbf{x}_{it}; \boldsymbol{\beta})$$

Then, we approximate $g_t(.)$ with a $\mathcal{P}^{\text{th}}$ order polynomial in its argument:

$$\begin{aligned}\omega_{it}(\boldsymbol{\beta}) &= \Omega_{it-1}(\boldsymbol{\beta})'\rho(\boldsymbol{\beta}) + \xi_{it} \\ &= \sum_{p=0}^{\mathcal{P}} \rho_p \omega_{it-1}^p(\boldsymbol{\beta}) + \xi_{it}\end{aligned}$$

where we follow De Loecker and Warzynski (2012) and set $\mathcal{P} = 3$. Thus, the innovations to productivity can be constructed as a function of $\boldsymbol{\beta}$ through:

$$\xi_{it}(\boldsymbol{\beta}) = \omega_{it}(\boldsymbol{\beta}) - \Omega_{it-1}(\boldsymbol{\beta})'\widehat{\rho}(\boldsymbol{\beta})$$

The estimates $\widehat{\rho}(\boldsymbol{\beta}) = (\{\widehat{\rho}_p\}_{p=1}^{\mathcal{P}})'$ are simply obtained by running a least squares regression of $\Omega_{it-1}(\boldsymbol{\beta})$ on $\omega_{it}(\boldsymbol{\beta})$.

Step 3. GMM-IV estimation of $\boldsymbol{\beta}$.

By assumption, capital is predetermined at time t as a firm chooses it one period ahead. As a result, it is safe to assume that k_{it} is orthogonal to the innovation $\xi_{it}(\boldsymbol{\beta})$. Similarly, firms cannot observe the string of future innovations to their productivity. As a result, current input decisions (with the exception of investment in capital) must be orthogonal to shocks to their idiosyncratic productivity in the future. Define the instrument $\mathbf{z}_{it} \in \mathbb{R}^Z$ as the vector that contains one-period lagged values of every polynomial term containing ℓ_{it}, m_{it} and e_{it} in the production technology $f(\mathbf{x}_{it}; \boldsymbol{\beta})$ but with capital preserved at their current values k_{it}. Then, we define the following system of moment conditions to identify $\boldsymbol{\beta} \in \mathbb{R}^Z$:

$$\mathbb{E}\left(\xi_{it}(\boldsymbol{\beta})\mathbf{z}_{it}\right) = \mathbf{0}_{Z\times 1} \tag{25}$$

By construction, this system of equations defines a set of exogeneity conditions. Lagged inputs are used to instrument for current period inputs. To validate this identification strategy, we need to argue that the moment conditions in (25) also satisfy rank conditions. Our focus lies on material inputs, so we will pay particular attention for this specific input. For lagged material inputs to be a valid instrument for current material inputs, m_{it} and m_{it-1} need to be correlated. A sufficient condition would be that input prices for material inputs are persistent over time. In fact, Atalay (2014) finds empirical evidence for this using data from the Census of Manufactures.

To obtain $\boldsymbol{\beta}$, we rely on the minimization of a quadratic loss function which is standard in GMM estimation.[27] Thus, we get:

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^Z}{\arg\min} \sum_{m=1}^{Z}\left(\sum_{i=1}^{N}\sum_{t=1}^{T}\xi_{it}(\boldsymbol{\beta})z_{it}^{m}\right)^2$$

Constructing markups after obtaining estimates $\widehat{\boldsymbol{\beta}}$. In general, output elasticities with respect to material inputs can depend on the level of *all* inputs; be it flexible or predetermined. This implies that $\theta_{it}^{M} = \theta_{M}^{j(i)}(\mathbf{x}_{it};\boldsymbol{\beta})$. Following the estimation procedure by De Loecker and Warzynski (2012), we can furthermore correct for measurement error ϵ_{it} in logged output. This is particularly important for data in the ASM and CM. Output prices are not available at the firm level, so output levels are obtained by deflating revenues adjusted for inventories. Unfortunately, the deflators used in the NBER-CES Manufacturing database only has deflators available at the industry level. This causes an unavoidable bias in measuring real output.

However, De Loecker and Warzynski (2012) mention that some concern of this bias can be taken care of with the correction term ϵ_{it}. By construction, any unobserved variation in output prices orthogonal to a firm's inputs will be absorbed by the measurement error correction term. In addition, if pricing decisions are correlated with a plant's productivity, then this specific variation will be controlled for as well through the use of a proxy for productivity. Then, markups are constructed as:

$$\begin{aligned}\widehat{\mu}_{it} &= \widehat{\theta}_{it}^{M}\left(\frac{\texttt{vm}_{it}}{\texttt{tvs}_{it}/\widehat{\epsilon}_{it}}\right)^{-1}\\ &= \theta_{M}^{j(i)}(\mathbf{x}_{it};\widehat{\boldsymbol{\beta}})\left(\frac{\texttt{vm}_{it}}{\texttt{tvs}_{it}/\exp(\widehat{\epsilon}_{it})}\right)^{-1}\end{aligned} \tag{26}$$

where $\texttt{vm}_{it}$ and $\texttt{tvs}_{it}$ denote a plant i's total expenditure on intermediate inputs and total value of shipments in year t. Production technologies do not differ over time but are allowed to vary across industries by

[27] By construction, the number of parameters in $\boldsymbol{\beta}$ is equal to the amount of identifying moments. This case of "just identification" renders the specification of a weighting matrix useless.

assumption **A6**.[28]

To construct output elasticities explicitly, we need to take a stance on the production function. In the following, we demonstrate how to obtain output elasticities in the case of Cobb-Douglas and translog production. Our preferred specification assumes that production is translog. This has two reasons. First, the translog specification is a second-order log approximation to *any* arbitrary production function. In fact, the Cobb-Douglas setup is nested within our translog specification. Second, output elasticities are allowed to vary with the level of any input under the translog specification. This implies that markups and markdowns have two sources of time variation: time-varying output elasticities and input revenue shares.

COBB DOUGLAS PRODUCTION. Assumption **A6** under Cobb-Douglas production immediately implies that $\theta_{it}^M = \beta_M^{j(i)}$ where $j(i)$ denotes the industry of some plant i. To implement the estimation procedure by De Loecker and Warzynski (2012), we have:

$$f(\mathbf{x}_{it};\boldsymbol{\beta}) = \beta_K k_{it} + \beta_L \ell_{it} + \beta_M m_{it} + \beta_E e_{it}$$

Furthermore, the instrument vector boils down to $\mathbf{z}_{it} = (k_{it}, \ell_{it-1}, m_{it-1}, e_{it-1})'$ and markdowns are empirically implemented as:

$$\widehat{\nu}_{it}^{\text{CD}} = \widehat{\beta}_L^{j(i)} \left(\frac{\texttt{sw}_{it}}{\texttt{tvs}_{it}}\right)^{-1} \left[\widehat{\beta}_M^{j(i)} \left(\frac{\texttt{vm}_{it}}{\texttt{tvs}_{it}/\exp(\widehat{\epsilon}_{it})}\right)^{-1}\right]^{-1}$$

where $\texttt{sw}_{it}$ denotes a plant i's total wage bill in year t. Note that $\boldsymbol{\beta} \in \mathbb{R}_+^4$ is estimated for each industry group j.

TRANSLOG PRODUCTION. Assumption **A6** under translog production implies:

$$\begin{aligned} f(\mathbf{x}_{it};\boldsymbol{\beta}) = {} & \beta_K k_{it} + \beta_L \ell_{it} + \beta_M m_{it} + \beta_E e_{it} \\ & + \beta_{KL} k_{it}\ell_{it} + \beta_{KM} k_{it} m_{it} + \beta_{KE} k_{it} e_{it} + \beta_{LM} \ell_{it} m_{it} + \beta_{LE} \ell_{it} e_{it} + \beta_{ME} m_{it} e_{it} \\ & + \beta_{KK} k_{it}^2 + \beta_{LL} \ell_{it}^2 + \beta_{MM} m_{it}^2 + \beta_{EE} e_{it}^2 \end{aligned}$$

Assuming that capital is chosen one period ahead, the instrument vector becomes:

$$\begin{aligned} \mathbf{z}_{it} = \Big(& k_{it}, \ell_{it-1}, m_{it-1}, e_{it-1}, k_{it}\ell_{it-1}, k_{it} m_{it-1}, k_{it} e_{it-1}, \ell_{it-1} m_{it-1}, \ell_{it-1} e_{it-1}, m_{it-1} e_{it-1}, \\ & k_{it}^2, \ell_{it-1}^2, m_{it-1}^2, e_{it-1}^2 \Big)' \end{aligned}$$

[28] Note that this assumption can be relaxed by estimating time-varying Cobb-Douglas parameters. This is easily done by restricting the estimation sample to a single cross-section in a given year. Theoretically, this should be possible for the translog case as well, but the amount of cross-sectional variation in a given industry-year cell might not be sufficient to identify all parameters properly.

where $\boldsymbol{\beta} \in \mathbb{R}^{14}$ is estimated for each industry j. Note that the number of parameters increases exponentially whenever more inputs are considered.[29] Markdowns are then empirically implemented through:

$$\widehat{\nu}_{it}^{\text{TL}} = \widehat{\theta}_{\ell}^{j(i)}(\mathbf{x}_{it}; \widehat{\boldsymbol{\beta}}) \left(\frac{\texttt{sw}_{it}}{\texttt{tvs}_{it}}\right)^{-1} \left[\widehat{\theta}_{M}^{j(i)}(\mathbf{x}_{it}; \widehat{\boldsymbol{\beta}}) \left(\frac{\texttt{vm}_{it}}{\texttt{tvs}_{it}/\exp(\widehat{\epsilon}_{it})}\right)^{-1}\right]^{-1}$$

s.t.

$$\begin{aligned}
\widehat{\theta}_{\ell}^{j(i)}(\mathbf{x}_{it}; \widehat{\boldsymbol{\beta}}) &= \widehat{\beta}_{L}^{j(i)} + \widehat{\beta}_{KL}^{j(i)} k_{it} + \widehat{\beta}_{LM}^{j(i)} m_{it} + \widehat{\beta}_{LE}^{j(i)} e_{it} + 2\widehat{\beta}_{LL}^{j(i)} \ell_{it} \\
\widehat{\theta}_{M}^{j(i)}(\mathbf{x}_{it}; \widehat{\boldsymbol{\beta}}) &= \widehat{\beta}_{M}^{j(i)} + \widehat{\beta}_{KM}^{j(i)} k_{it} + \widehat{\beta}_{LM}^{j(i)} \ell_{it} + \widehat{\beta}_{ME}^{j(i)} e_{it} + 2\widehat{\beta}_{MM}^{j(i)} m_{it}
\end{aligned}$$

A.3 Identification

Under the above described "production function" approach of obtaining markdowns, we rely on proxy variable methods to obtain output elasticities. While the induced moment conditions are easily derived and understood, Gandhi, Navarro and Rivers (2017) emphasize that point identification is not achieved when applying the methodology by De Loecker and Warzynski (2012), for example. To address this criticism, we apply the solution suggested in Flynn, Gandhi and Traina (2019). They show that the non-identification problem can be resolved whenever the returns to scale of the production function is ex-ante specified. Similar to their work, we show the robustness of our markdown estimates whenever we impose a constant returns to scale restriction.[30] This seems reasonable since a substantial body of previous work (e.g., Basu and Fernald, 1997; Syverson, 2004a; Syverson, 2004b) has shown that constant returns to scale is a good approximation for manufacturing plants.

In the following, we will briefly describe how our estimation procedure is adjusted (for the translog case) when imposing constant returns to scale. In fact, this requires minor adjustments only. Steps 1 and 2 are unchanged whereas we only need to add some moment conditions to step 3. To do so, define a firm's returns to scale as follows:

$$\Sigma_{it}(\boldsymbol{\beta}) = \sum_{\iota \in \{k,\ell,m,e\}} \frac{\partial f(\mathbf{x}_{it}; \boldsymbol{\beta})}{\partial \iota_{it}} \tag{27}$$

Also, define the vector $\boldsymbol{\chi}_{it} = (1, \mathbf{x}_{it}')' \in \mathbb{R}^{K+1}$, then the new set of moment conditions can be compactly written as:

$$\mathbb{E}\begin{pmatrix} \xi_{it}(\boldsymbol{\beta})\mathbf{z}_{it} \\ \Sigma_{it}(\boldsymbol{\beta})\boldsymbol{\chi}_{it} \end{pmatrix} = \mathbf{0}_{(Z+K+1)\times 1} \tag{28}$$

[29] With a translog production function with K inputs, there are K linear terms, K quadratic components and $\binom{K}{2}$ unique input pairs. Thus, there are a total of $2K + \binom{K}{2} = \frac{K(K+3)}{2}$ terms.

[30] We draw similar conclusions whenever we allow for slight deviations from constant returns to scale.

A.4 Material inputs as a flexible factor of production

The identification of markdowns (and markups) hinges upon the assumption that material inputs are flexible. While this is a standard assumption in the literature (see, for example, Basu (1995) and De Loecker and Warzynski (2012)), it is not easy to validate this assumption in the data directly.[31] This has caused some papers to move away from material inputs and use other inputs instead.

Kim (2017) also relies on the insight by Hall (1988) to identify markups but rather than using material inputs he advocates that energy inputs are better suited for the estimation of markups. While static or dynamic adjustment costs are not present for material inputs, he argues that forces of monopsony power are more present in markets for material inputs due to buyer-supplier networks. As a result, this would violate assumption **A3**. Energy inputs however are less prone to this problem as prices for energy tend to be regulated or supplied by government agencies and/or large organizations.

However, Davis et al. (2013) find that plant-level differences within manufacturing industries in energy purchases account for a substantial fraction of overall price dispersion. At least one third of cross-sectional dispersion in log electricity prices is due to variation between purchase deciles (or quantiles).[32] Furthermore, price gaps between larger and smaller purchases are enormous; even when controlling for plant location and/or electric utility provider fixed effects. This seems to contradict the "no monopsony" condition for energy embodied in assumption **A3**. Furthermore, revenue shares of energy inputs are extremely low in U.S. manufacturing implying that a small amount of measurement error in energy inputs can create substantial distortions in our markup estimates. This is less problematic for material inputs since a large fraction of revenue is covered by the latter input. We view these results as evidence speaking in favor of material inputs.

Another reason why material inputs are preferable is that the required rank conditions of the GMM-IV estimation procedure by De Loecker and Warzynski (2012) can be explicitly tested in the data. It is sufficient to satisfy the rank condition whenever prices for material inputs are persistent. Evidence for this observation from data similar to this paper is documented in Atalay (2014) providing additional justification for the validity of the estimation procedure used in this paper.

DIRECT EVIDENCE FROM THE COMMODITY FLOW SURVEY. In this section, we use data from the Commodity Flow Survey (CFS) to construct empirical price-quantity schedules for a subset of material inputs.

To do so, it is first necessary to establish what constitutes "materials". We follow the definition of the Census Bureau and construct our primary measure for material inputs as the deflated sum of expenditures on materials and parts, resales and contract work. The previously mentioned deflator, which is constructed at

[31] Morlacco (2019) uses transaction-level data from French manufactures to show that there is substantial market power in foreign intermediate inputs. To reach this conclude however, it is assumed that domestic intermediate inputs are perfectly competitive.

[32] According to Davis et al. (2013), this fraction was as high as 75 percent in 1963 but dropped to 30 percent by 1978.

the 4-digit SIC level, is taken from the NBER-CES Manufacturing Database and specifically constructed for materials.

To construct price-quantity schedules, it is necessary to observe prices and quantities for material inputs separately. Unfortunately, the ASM/CM only has information on expenditures. Therefore, we use transaction-level data from the CFS which contains a representative sample of transactions carried out by establishments in mining, manufacturing, wholesale trade, and select retail and service industries. A big advantage of this micro-level data is that expenditures and quantities can be observed separately which allows us to construct prices. Recall that we are only interested in transactions that deal with material inputs. To identify these particular transactions, we use two sources of information. First, we use the Standard Classification of Transported Goods (SCTG) codes that are available in the CFS. These are essentially commodity codes. Second, we identify what products are labeled as "material inputs" in the ASM/CM. According to the Census Bureau, material inputs contain materials, parts, containers and supplies. In section 16A1 of form MA-10000, we observe that each of these groups contain a few product categories which are outlined in the table below.

Table VI: Description of section 16A1 in form MA-10000 of Annual Survey of Manufactures (ASM).

MATERIALS		PARTS	CONTAINERS	SUPPLIES	
Lumber	Cement	Pumps	Pails	Bolts, screw and nuts	Cleaning supplies
Plywood	Clay	Wheels	Drums and barrels	Drills, tools, dies, jigs and fixtures which are charged to current accounts	Stationary and office supplies
Paper	Glass	Bearings	Tubes	Welding rods, electrodes and acetylene	First aid and safety supplies
Resins	Steel sheet	Engines	Boxes and bags	Lubricating oils	Dunnage
Sulfuric acid	Steel scrap	Gears	Crates		Water
Alcohols	Copper rods	Motors			
Rubber	Iron castings	Hardware			
Coking coal	Metal stampings	Compressors			
Crude petroleum	Wire				

For each entry in the above table, we then manually match what SCTG codes are contained by it. For example, lumber is associated with SCTG codes 25020 (Logs for lumber), 26211 (Lumber, treated) and 26212 (Lumber, untreated). This procedure allowed us to identify 67 SCTG codes that were classified as either materials, parts, containers or supplies. Our final sample only considers domestic transactions that contain one of these 67 SCTG codes in 1993, 1997, 2002, 2007 or 2012.

Following Davis et al. (2013), we construct empirical price-quantity schedules by year. We do this for raw prices and quantities, and while controlling for commodity fixed effects to deal with product quality and other dimensions of heterogeneity. In the end, we find that the slopes of these empirical price-quantity schedules are comparable to those found in Davis et al. (2013).

A.5 Measures of compensation

In our baseline estimation procedure, we measure a plant's total wage bill (or "payroll") deflated by the NBER-CES shipments deflator. Following the instructions of form MA-10000, payroll is an overall measure of wages and salaries paid to a plant's employee(s). An employee is defined according to Internal Revenue Service Form 941, Employer's Quarterly Federal Tax Return. This includes:

- All persons on paid sick leave, paid holidays, and paid vacation during these pay periods
- Officers at this establishment, if a corporation
- Spread on stock options that are taxable to employees as wages

An employer's wage bill is defined as its payroll before deductions excluding an employer's cost for fringe benefits. In particular, it includes:

- Employee's Social Security contributions, withholding taxes, group insurance premiums, union dues, and savings bonds
- In gross earnings: commissions, dismissal pay, paid bonuses, employee contributions to pension plans such as 401(k), vacation and sick leave pay, and the cash equivalent of compensation paid in kind
- Spread on stock options that are taxable to employees as wages
- Salaries of officers of this establishment, if a corporation
- Paid holiday, personal, funeral, jury duty, military and family leave
- Non-production bonuses
  - Cash profit-sharing
  - Employee recognition
  - End-of-year
  - Holiday
  - Payment in lieu of benefits - Referral
  - Other

By construction, the wage bill does not include benefits. Fortunately, the ASM/CM does include a measure of these benefits from 2002 onward. Benefits cover health insurance, pension plans and other employer paid benefits. The latter includes legally-required benefits (e.g., Social Security, workers' compensation insurance, unemployment tax, state disability insurance programs, Medicare), benefits for life insurance, "quality of life" benefits (e.g., childcare assistance, subsidized commuting, etc.), employer contributions to pre-tax benefit accounts (e.g., health savings accounts), education assistance, and other benefits. Our results on markdowns are not qualitatively changed whenever we use a measure for labor that includes benefits.

B Derivation on markdowns

B.1 Aggregation of micro-level markdowns

Proof of proposition 2. Whenever assumptions **A1** – **A5** are satisfied and assumptions **A2** – **A5** apply to material inputs, then we showed in lemma 1 that markups can be characterized as:

$$\begin{aligned}\mu_{it} &= \frac{\theta_{it}^M}{\alpha_{it}^M} \\ &= \theta_{it}^M \cdot \frac{p_{it}y_{it}}{P_t^M m_{it}} \qquad (29)\end{aligned}$$

Similar to Edmond, Midrigan and Xu (2019), we define the **aggregate markup** as the wedge between the aggregate output elasticity of some flexible input and its revenue share. Under the assumption of material inputs being flexible, equation (26) also holds in the aggregate, i.e. we have:

$$\mathcal{M}_t \equiv \theta_t^M \cdot \frac{P_t Y_t}{P_t^M M_t} \qquad (30)$$

where we dropped the indices for local markets (j, ℓ) for simplicity. Substituting out the price for material inputs P_t^M from (27) into (26), we obtain:

$$\mu_{it} = \frac{\theta_{it}^M}{\theta_t^M} \cdot \frac{p_{it}y_{it}}{P_t Y_t} \cdot \frac{M_t}{m_{it}} \cdot \mathcal{M}_t$$

Then, we sum across firms and rearrange to derive the aggregate markup:

$$\mathcal{M}_t = \left(\sum_{i \in F_t} \frac{\theta_{it}^M}{\theta_t^M} \cdot s_{it} \cdot \mu_{it}^{-1} \right)^{-1} \qquad (31)$$

where $s_{it} \equiv \frac{p_{it}y_{it}}{P_t Y_t}$ denotes a firm i's revenue share relative to the aggregate and we used the definition for aggregate materials $M_t = \sum_{i \in F_t} m_{it}$. Whenever production technologies are Cobb-Douglas, we have $\theta_{it}^M = \theta_t^M$ for each $i \in F_t$. Then, the aggregate markup is simply a revenue-weighted harmonic average of firm-level markups.

We use a similar insight to derive the **aggregate markdown** $\mathcal{V}_t$. In the absence of adjustment costs for labor, the wedge between the output elasticity of labor and its revenue share for a firm i must reflect market power in either output or labor markets. We showed this explicitly in proposition 3. Therefore, we have:

$$\nu_{it}\mu_{it} = \theta_{it}^L \cdot \frac{p_{it}y_{it}}{w_{it}\ell_{it}} \qquad (32)$$

Rearranging for a firm i's wage bill and summing across firms, it follows that:

$$\begin{aligned}\sum_{i \in F_t} w_{it}\ell_{it} &= w_t L_t \\ &= P_t Y_t \cdot \sum_{i \in F_t} \theta_{it}^L \cdot s_{it} \cdot (\mu_{it}\nu_{it})^{-1}\end{aligned}$$

where the first equality follows from definition of the aggregate wage bill. We define the aggregate markdown $\mathcal{V}_t$ as that part of the wedge between the aggregate output elasticity of labor and the aggregate labor share that is not due to markups. Then, by definition, we have:

$$\mathcal{V}_t \cdot \mathcal{M}_t = \theta_t^L \cdot \frac{P_t Y_t}{w_t L_t} \tag{33}$$

Using our previous results, we then get:

$$\begin{aligned}\mathcal{V}_t \cdot \mathcal{M}_t &= \theta_t^L \cdot \left(\sum_{i \in F_t} \theta_{it}^L \cdot s_{it} \cdot (\mu_{it}\nu_{it})^{-1} \right)^{-1} \\ &= \left(\sum_{i \in F_t} \frac{\theta_{it}^L}{\theta_t^L} \cdot s_{it} \cdot (\mu_{it}\nu_{it})^{-1} \right)^{-1}\end{aligned}$$

Apply expression (28) for the aggregate markup and we obtain an expression for the aggregate markdown:

$$\mathcal{V}_t = \frac{\left(\sum_{i \in F_t} \frac{\theta_{it}^L}{\theta_t^L} \cdot s_{it} \cdot (\mu_{it}\nu_{it})^{-1} \right)^{-1}}{\left(\sum_{i \in F_t} \frac{\theta_{it}^M}{\theta_t^M} \cdot s_{it} \cdot \mu_{it}^{-1} \right)^{-1}} \tag{34}$$

A special case is whenever each firm i has a Cobb-Douglas technology. Then, we get:

$$\mathcal{V}_t = \frac{\left(\sum_{i \in F_t} s_{it} \cdot (\mu_{it}\nu_{it})^{-1} \right)^{-1}}{\left(\sum_{i \in F_t} s_{it} \cdot \mu_{it}^{-1} \right)^{-1}} \tag{35}$$

which amounts to a ratio of sales-weighted harmonic averages. □

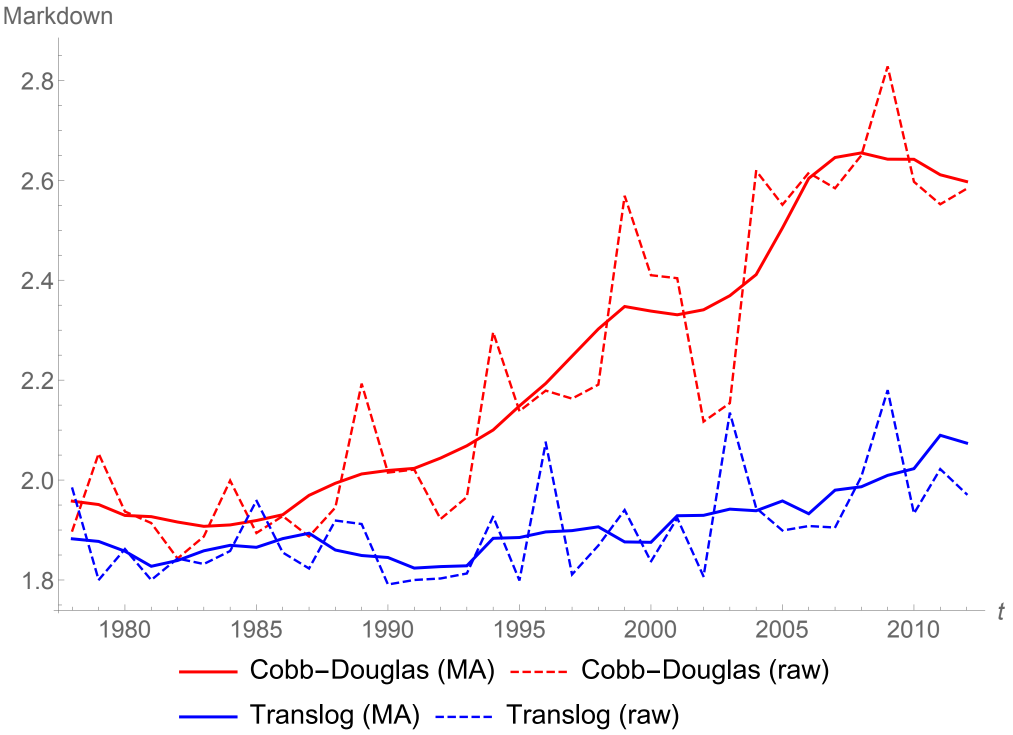
B.2 Secular trend in aggregate markdowns: Cobb-Douglas versus translog

In our baseline estimates, we specified production functions to be translog. By construction, the translog specification allows output elasticities to vary with the level of inputs. As a result, these output elasticities can vary over time as well. Under a Cobb-Douglas specification, output elasticities are constant and markdowns can only vary over time due to changes in revenue shares. In the following, we show that allowing for time-varying output elasticities is important for several measures of the aggregate markdown.

We start with the aggregate markdown measure that is based on De Loecker, Eeckhout and Unger (2018). In particular, we will show that distinguishing between Cobb-Douglas and translog markdowns at the micro-levels matters greatly for the measure $\mathcal{V}_t^{\text{dLEU}}$.

Figure 7 shows that the measure $\mathcal{V}_t^{\text{dLEU}}$ is increasing over time for both Cobb-Douglas and translog markdowns. However, the differences between Cobb-Douglas and translog are quantitatively quite stark. Indeed, under the assumption of Cobb-Douglas production functions, aggregate markdowns increased from 1.96 to 2.60 (relative increase of 33 percent) over the period 1976 – 2014. The results for translog technologies show an increase from 1.89 to 2.07 over the same period, a much weaker uptick of 10 percent. These differences underline that Cobb-Douglas specifications can be quite restrictive. By construction, the Cobb-Douglas specification assumes that output elasticities are constant and, hence, ignores any time variation in a plant's output elasticities. Conversely, a translog specification allows precisely for this. Our results favor the translog specification since they indicate that this time variation is quantitatively important.

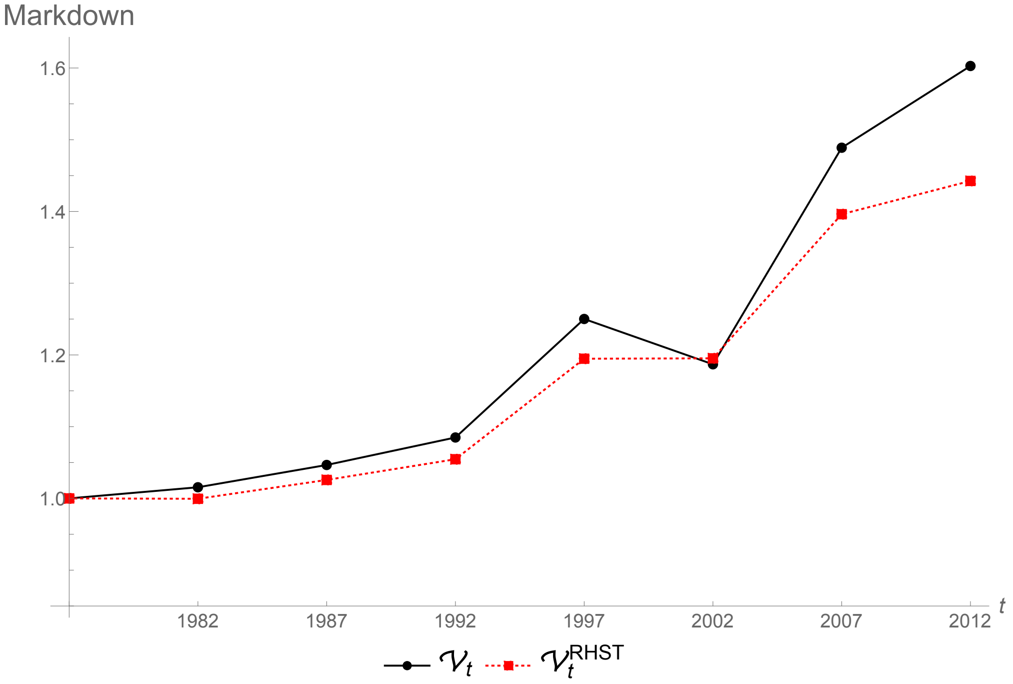
Figure 7: Time evolution of employment-weighted markdowns across U.S. manufacturing plants from 1976 to 2014.



Markdowns are constructed under the assumption of translog (blue) or Cobb-Douglas (red) production and aggregated according to expressions equation (14). Source: authors' own calculations from ASM/CM data in 1976–2014.

The difference between Cobb-Douglas and translog markdowns has an even greater impact for the measures $\mathcal{V}_t^{\text{RHST}}$ and $\mathcal{V}_t$. While these measures are decreasing over time (at least before 2002) under a translog specification, the opposite is true whenever markdowns are estimated under Cobb-Douglas technologies.

Figure 8: Time evolution of employment-weighted markdowns across U.S. manufacturing plants from 1977 to 2012.



Markdowns are constructed under the assumption of Cobb-Douglas production and aggregated according to expressions equation (13) and equation (15), respectively. All measures are normalized relative to their initial value in 1977. Source: authors' own calculations from quinquennial CM data from 1977–2012.

C Labor adjustment costs

In this appendix, we show that the wedge between the marginal revenue product of labor and the wage is no longer reflective of only labor market power whenever labor adjustment costs are present. This is not a trivial result since a firm's profit maximization problem becomes dynamic when labor is subject to costly adjustments. Intuitively, this is because labor adjustment costs depend on the level of labor in the previous period. If these adjustment costs take a quadratic form however, it is possible to "correct" our initial estimates for markdowns. When we apply these correction terms to our estimates, we obtain measures for markdowns that are only reflective of monopsony forces and not of labor adjustment costs. In the end, we find that these correction terms are quantitatively small.

The proposition below shows that labor adjustment costs can also drive a wedge between marginal revenue products of labor and wages. Nevertheless, we can identify the "monopsony" component whenever these adjustment costs take a quadratic form.

PROPOSITION 4. Let $\mathbf{z}$ denote a firm's set of stochastic state variables and suppose revenue, labor adjustment cost and wage schedule functions are differentiable. Then, a firm's wedge between its MRPL and wage satisfies:

$$\frac{R'(\ell^*)}{w(\ell^*)} = \left(\varepsilon_S^{-1} + 1\right) + \mathcal{A}(\ell^*, \ell_{-1})$$

where $\mathcal{A}(\ell^*, \ell_{-1})$ equals zero whenever labor adjustment costs are absent. If, in addition, a firm is subject to convex labor adjustment costs of the form $\Phi(\ell, \ell_{-1}) = \frac{\gamma}{2}\left(\frac{\ell-\ell_{-1}}{\ell_{-1}}\right)^2$ for $\gamma \geq 0$ and it discounts future profits at the rate $\beta \in [0, 1]$, then a firm's monopsony power can be characterized as:

$$\varepsilon_S^{-1} + 1 = \frac{\frac{R'(\ell^*)}{w(\ell^*)} - \gamma \cdot \left(g_\ell(1+g_\ell) - \beta \mathbb{E}_{\mathbf{z}'}\left[g_{\ell'}(1+g_{\ell'})(1+g_{\mathrm{sw}'})|\mathbf{z}\right]\right)}{1 + \frac{\gamma}{2} g_\ell^2} \tag{36}$$

where g_ℓ, $g_{\ell'}$ and $g_{\mathrm{sw}'}$ denote current and future labor growth, and future wage bill growth, respectively.

Proof. We will consider environments in which revenue, labor adjustment costs and wage schedules are continuously differentiable (at least in labor). Furthermore, we will restrict our attention to convex adjustment costs in labor, but we do allow for dynamic considerations (i.e., adjustment costs in labor are allowed to depend on the stock of labor in the previous period; denoted by ℓ_{-1}). Then, consider a firm's *dynamic* profit maximization problem:

$$v(\ell_{-1}; \mathbf{z}) = \max_{\ell \geq 0} R(\ell; \mathbf{z}) - w(\ell) \cdot \ell - w(\ell) \cdot \Phi(\ell, \ell_{-1}) + \beta \cdot \mathbb{E}_{\mathbf{z}'}\left[v(\ell; \mathbf{z}')|\mathbf{z}\right] \tag{37}$$

where $\Phi(\ell, \ell_{-1})$ denotes a firm's adjustment cost (in real terms) whenever it wants to change its stock of labor to $\ell \neq \ell_{-1}$ and $\beta \in [0, 1]$ is its discount factor. We will assume that the adjustment cost function is homogeneous of degree one and continuously differentiable in both arguments. Furthermore, we have that $\Phi(\ell, \ell_{-1}) > 0$ for $\ell \neq \ell_{-1}$ and zero otherwise. Similar to before, we denote the revenue function by $R(\ell; \mathbf{z}) \equiv \mathrm{rev}(\ell; \mathbf{x}^*_{-\ell}(\ell), \mathbf{z})$ where $\mathbf{z}$ denotes a firm's (possibly stochastic) state variable, e.g. productivity. Given this setup, a firm's optimal choice is characterized by its first order condition:

$$\begin{aligned} R'(\ell) &= w'(\ell)\ell + w(\ell) + w(\ell) \cdot \Phi_1(\ell, \ell_{-1}) + w'(\ell) \cdot \Phi(\ell, \ell_{-1}) - \beta \cdot \mathbb{E}_{\mathbf{z}'}\left[v'(\ell)|\mathbf{z}\right] \\ &= w'(\ell)\ell + w(\ell) + w(\ell) \cdot \Phi_1(\ell, \ell_{-1}) + w'(\ell) \cdot \Phi(\ell, \ell_{-1}) + \beta \cdot \mathbb{E}_{\mathbf{z}'}\left[\Phi_2(\ell', \ell) w(\ell')|\mathbf{z}\right] \end{aligned}$$

where we applied the envelope theorem in the last equality. This can be rearranged to end up with an expression for a firm's markdown:

$$\begin{aligned} \nu &\equiv \frac{R'(\ell)}{w(\ell)} \\ &= \varepsilon_S^{-1} + 1 + \Phi_1(\ell, \ell_{-1}) + \frac{\Phi(\ell, \ell_{-1})}{\ell}\varepsilon_S^{-1} + \beta \cdot \mathbb{E}_{\mathbf{z}'}\left[\Phi_2(\ell', \ell)\frac{w(\ell')}{w(\ell)}\middle| \mathbf{z}\right] \\ &\equiv \varepsilon_S^{-1} + 1 + \mathcal{A}(\ell, \ell_{-1}) \end{aligned} \tag{38}$$

where $\mathcal{A}(\ell, \ell_{-1})$ reflects a firm's expected continuation value of adjustment cost relative to its wage level.

Without specifying the shape of the real labor adjustment cost function further, it is hard to assess the magnitude of the bias (i.e., $\mathcal{A}(\ell, \ell_{-1})$) that we are dealing with. For illustrative purposes, we use a commonly specified labor adjustment cost function $\Phi(\ell, \ell_{-1}) = \frac{\gamma}{2}\ell\left(\frac{\ell-\ell_{-1}}{\ell_{-1}}\right)^2$ (Hall, 2004; Cooper, Haltiwanger and Willis, 2007). Given this specification and after some algebra, we can simplify equation (38) to:

$$\nu = \left(1 + \frac{\gamma}{2}g_\ell^2\right)(\varepsilon_S^{-1} + 1) + \gamma g_\ell(1 + g_\ell) - \beta\gamma\mathbb{E}_{\mathbf{z}'}\left[g_{\ell'}(1 + g_{\ell'})(1 + g_{\mathtt{SW}'})|\mathbf{z}\right] \tag{39}$$

where we defined labor growth rates as $g_\ell = \frac{\ell-\ell_{-1}}{\ell_{-1}}$ and $g_{\ell'} = \frac{\ell'-\ell}{\ell}$, respectively. Furthermore, we have a firm's future growth rate in its wage bill which equals $g_{\mathtt{SW}'} = \frac{w(\ell')\ell'}{w(\ell)\ell} - 1$. If our estimates for markdowns do not only reflect monopsony, then we can obtain "unbiased" estimates for labor market power (i.e., percentage wedges between marginal revenue products of labor and wages corrected for labor adjustment costs as reflected by $\varepsilon_S^{-1} + 1$ alone) by using equation (39) instead. To do so, we solve for $\varepsilon_S^{-1} + 1$ and obtain:

$$\varepsilon_S^{-1} + 1 = \frac{\frac{R'(\ell^*)}{w(\ell^*)} - \gamma \cdot \left(g_\ell(1 + g_\ell) - \beta\mathbb{E}_{\mathbf{z}'}\left[g_{\ell'}(1 + g_{\ell'})(1 + g_{\mathtt{SW}'})|\mathbf{z}\right]\right)}{1 + \frac{\gamma}{2}g_\ell^2}$$

which is exactly what we wanted to show. □

We apply the above proposition by substituting out expected growth rates with their realized counterparts. In particular, our estimates for markdowns $\widehat{\nu}$ can be adjusted for labor adjustment costs as follows:

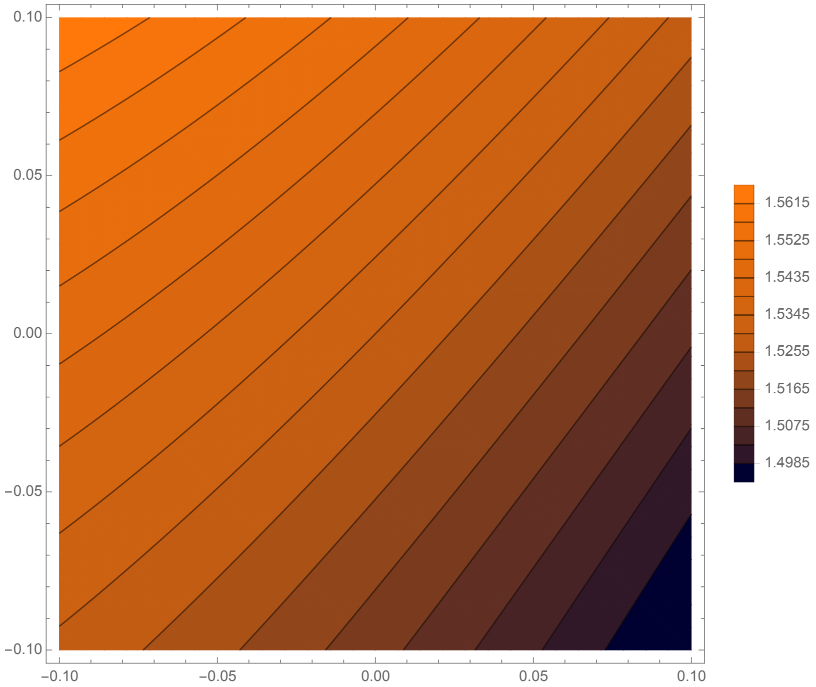
$$\frac{\varepsilon_S + 1}{\varepsilon_S} = \frac{\widehat{\nu} - \gamma \cdot \left[g_\ell(1 + g_\ell) - \beta g_{\ell'}(1 + g_{\ell'})(1 + g_{\mathtt{SW}'})\right]}{1 + \frac{\gamma}{2}g_\ell^2} \tag{40}$$

The proposition above shows that the wedge between a firm's MRPL and the wage it pays its workers no longer only reflects monopsony power in the presence of convex labor adjustment costs. In other words, labor adjustment costs can also drive a wedge between MRPL and wages. Hence, one could be worried that our measured markdowns do not only reflect monopsony forces but also capture labor adjustment costs.

If labor adjustment costs are quadratic, then the second part of the above proposition demonstrates that we can correct our measured markdowns such that they only reflect forces of monopsony power. This can be done if we observe a plant's growth in labor and its wage bill, and know the parameters β and γ. Obviously, quadratic adjustment costs are not without loss of generality, but it is a specification that is often employed (see Hall, 2004; Cooper, Haltiwanger and Willis, 2007). Another advantage of this functional form is that is governed by only one parameter. Obviously, we are back to our baseline in the absence of adjustment costs when $\gamma = 0$ holds as can be seen from equation (36).

To be conservative, we choose the highest estimate for γ in Hall (2004) that is estimated with reasonable

Figure 9: Correcting markdowns from convex labor adjustment costs.



(a) Absolute labor growth rates bounded by 10 percent

Wage bill growth $g_{\mathrm{sw}'}$ is set at 2.19 percent which is the average level of wage bill growth in U.S. manufacturing from 1987–2017 (BEA GDP by Industry accounts). Horizontal and vertical axes denote current and future labor growth g_ℓ and $g_{\ell'}$, respectively. The adjustment cost parameter γ is set at 0.185 (Hall, 2004).

precision. This results in $\gamma = 0.185$.[33] In figure 9, we set $\beta = 1$ and show that our measured markdowns only have to be adjusted by a maximum of 3.15 percent for a broad range of labor growth rates (varying from -10 to 10 percent). We conclude that labor adjustment costs only play a minor quantitative role and, hence, our baseline estimates must reflect labor market power.

D Labor market models with upward-sloping labor supply curves

D.1 Wage posting à la Burdett-Mortensen

For ease of notation, we drop a particular firm f's index. In the wage posting model of Burdett and Mortensen (1998), a firm's law of motion for its stock of labor is given by:

$$L_t = (1 - s(w_t))L_{t-1} + R(w_t) \tag{41}$$

where $s(\cdot)$ and $R(\cdot)$ denote the separation and recruiting functions, respectively. Note that these are allowed to explicitly depend on the posted wage. In a stationary setting, we must have $L_t = \frac{R(w_t)}{s(w_t)}$. Assuming that

[33] See the estimation results in table II of Hall (2004).

these functions are differentiable, it is straightforward to show that labor supply elasticities satisfy:

$$\varepsilon_S = \varepsilon_{Rw} - \varepsilon_{sw} > 0$$

where $\varepsilon_{Rw,t}$ and $\varepsilon_{sw,t}$ denote separation and recruiting elasticities, respectively. The above object is strictly positive since higher wages encourage hiring and lead to less separations, i.e. $\varepsilon_{Rw} > 0$ and $\varepsilon_{sw} < 0$.

Formally, the separation rate is induced by some exogenous job destruction process and poaching. In particular, we have $s(w) = \delta + \lambda(1 - F(w))$. Then, $-\varepsilon_{sw} = \lambda f(w) > 0$ follows directly from the fact that probability distribution functions are non-negative. Recall that the equilibrium wage distribution function has full support on $[0, \overline{w}]$ in the baseline framework of Burdett and Mortensen (1998). Furthermore, recruitment satisfies $R(w) = R^u + \lambda \cdot \int_0^w L(x)dF(x)$ where R^u is the stock of recruits from the pool of unemployment. Note that this does not vary across wage levels w since workers' values of unemployment are normalized to zero in Burdett and Mortensen (1998). Hence, unemployed workers accept any given offer. Given this structure, it is straightforward to derive that $\varepsilon_R = \lambda \cdot \frac{f(w)L(w)w}{R(w)} > 0$. While we only focused on the canonical model of Burdett and Mortensen (1998), upward-sloping labor supply curves are also present in more generalized settings such as Mortensen (2003) and Bontemps, Robin and Van den Berg (2001).

D.2 Additive Random Utility Models (ARUM)

In this section, we consider a class of additive random utility models as described in Chan, Kroft and Mourifie (2019). We do so because their setup nests a variety of labor market models which we will discuss below. There are K types indexed by k which each have a mass of m_k such that $\sum_{k=1}^{K} m_k = 1$. An individual worker i with type k (which is allowed to be multidimensional) is faced with the problem of choosing among a set of employers $\mathcal{J} = \{1, 2, \dots, J\}$. Its choice is informed by non-pecuniary benefits, wage compensation and some idiosyncratic term. A worker's outside option is denoted by "employer" 0. Its maximization problem is characterized by:

$$\max_{j \in \mathcal{J} \cup \{0\}} u_{kj} + w_{kj} + \varepsilon_{ij} = \max_{j \in \mathcal{J} \cup \{0\}} v_{kj} + \varepsilon_{ij}$$

The surplus function is defined as:

$$\mathcal{S}(\mathbf{v}_k) = \mathbb{E}\left[\max_{j \in \mathcal{J} \cup \{0\}} v_{kj} + \varepsilon_{ij}\right]$$

Then, Chan, Kroft and Mourifie (2019) characterize the labor supply function as:

$$\begin{aligned} L_{kj} &= m_k \cdot \Pr\left(v_{kj} + \varepsilon_{kj} \geq v_{kj'} + \varepsilon_{ij'}, \text{ for all } j' \in \mathcal{J} \cup \{0\}\right) \\ &= m_k \cdot \frac{\partial \mathcal{S}(\mathbf{v}_k)}{\partial v_{kj}} \end{aligned} \tag{42}$$

Chan, Kroft and Mourifie (2019) show that this object exists whenever ε_{ij} is independent of v_{kj} and is absolutely continuous with respect to the Lebesgue measure. Furthermore, the surplus function is convex in

$\mathbf{v}_k$ under those assumptions. Hence, labor supply schedules are non-decreasing. Therefore, we have:

$$\varepsilon_S^{kj} = \frac{m_k}{L_{kj}} \frac{\partial^2 \mathcal{S}(\mathbf{v}_k)}{\partial^2 v_{kj}} w_{kj} \geq 0$$

The generalized setting of Chan, Kroft and Mourifie (2019) is quite convenient as it nests the setups of Card et al. (2018) and Lamadon, Mogstad and Setzler (2019). This can be done by appropriately defining worker types and assuming that idiosyncratic shocks are drawn from an Extreme Value Type I distribution.

D.3 Monopsonistic competition

In the most simplistic setting, upward-sloping labor supply curves are purely generated through preferences, even in the absence of strategic complementarities across firms. For instance, this would be true in a setting in which a representative household supplies a bundle of differentiated labor $\mathbf{L}_t = \{L_{it}\}_{i=1}^K$ and has preferences over some composite consumption bundle C_t.

Suppose the household's preferences are summarized by some function $u(C_t, \mathbf{L}_t)$ that is continuously differentiable in its arguments. Then, the schedule of labor supply functions is determined by a system of non-linear equations consisting of $\frac{(K+1)K}{2} + 1$ equations. Intuitively, labor supply schedules are upward sloping whenever substitution effects dominate their income counterparts.

Horizontal job differentiation. Under this class of models, workers are heterogeneous in their preferences over non-wage characteristics of a job. A simple way to capture this idea is to assume that a worker's utility is increasing in wages and decreasing in distance to work. Then, wages act as a compensating differential. Examples are Bhaskar and To (1999) and Staiger, Spetz and Phibbs (2010) who adopt frameworks in the spirit of Salop (1979).[34]

Double-nested CES preferences (Atkeson-Burstein). Berger, Herkenhoff and Mongey (2019) consider a monopsonistic environment in the tradition of Atkeson and Burstein (2008). With some abuse of notation, preferences are characterized by:

$$u\left(C_t - \frac{1}{\varphi^{\frac{1}{\varphi}}} \frac{\mathbf{L}_t^{1+\frac{1}{\varphi}}}{1+\frac{1}{\varphi}}\right) \text{ with } \mathbf{L}_t = \left(\int_0^1 \mathbf{L}_{jt}^{\frac{\theta+1}{\theta}} dj\right)^{\frac{\theta}{\theta+1}} \text{ and } \mathbf{L}_{jt} = \left(\sum_{f=1}^{F_j} n_{fjt}^{\frac{\eta+1}{\eta}}\right)^{\frac{\eta}{\eta+1}}$$

Thus, preferences follow the GHH specification in consumption and labor whereas labor is a double-nested

[34] In particular, Staiger, Spetz and Phibbs (2010) assume that firms are uniformly distributed around a circle of measure one. Whenever the measure of firms N is fixed and workers' utility is increasing (decreasing) in their wage (distance to work), a firm i's labor supply function can be characterized as $L_i = \alpha + \tau^{-1}\left[w_i - \left(\frac{w_{i-1}+w_{i+1}}{2}\right)\right]$ where $\tau > 0$ denote travel costs (denoted in units of utility) per unit distance. Given this structure, we must have $\varepsilon_S > 0$.

CES composite. This gives rise to labor supply elasticities of the form:

$$\varepsilon_S = \frac{1}{\eta} + \left(\frac{1}{\theta} - \frac{1}{\eta}\right) \cdot s > 0$$

where $s \in [0, 1]$ is a firm's share of the industry's total payroll.

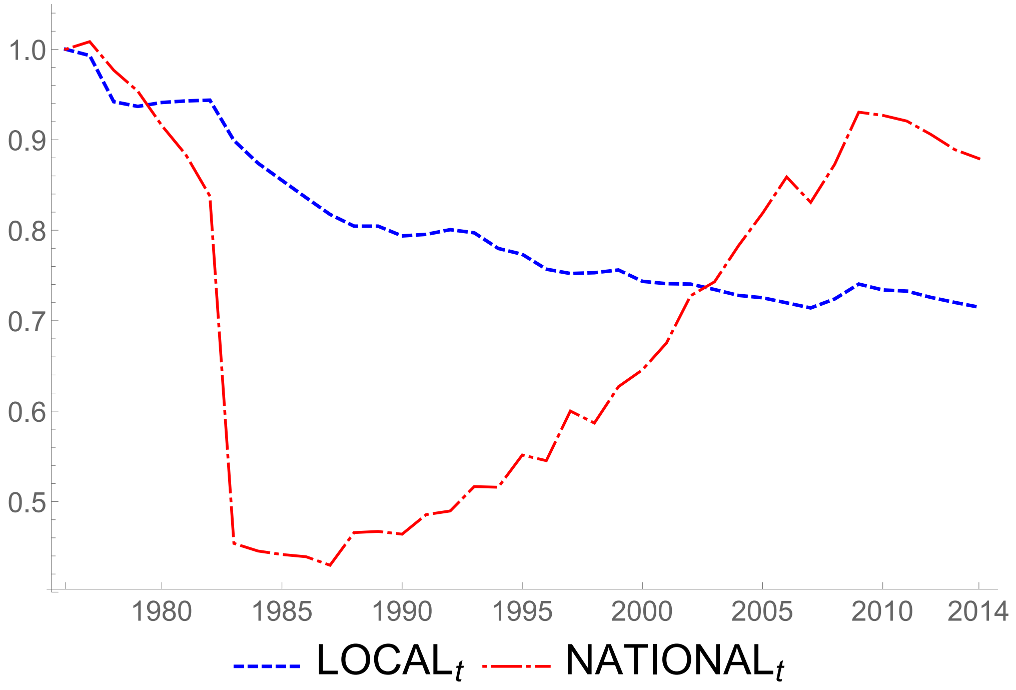
E National concentration

We construct **national** employment concentration, following Autor et al. (2017), as follows:

$$\begin{aligned} \text{NATIONAL}_t &= \sum_{j \in J} \omega_{jt} \text{HHI}_{jt} \\ &= \sum_{j \in J} \omega_{jt} \left[\sum_{f \in F_t(j)} \left(\frac{x_{ft}}{X_{F(j)t}} \right)^2 \right] \quad \text{s.t. } X_{F(j)t} = \sum_{f' \in F_t(j)} x_{f't} \end{aligned} \tag{43}$$

Hence, national concentration is a weighted average of industry-level HHIs. We implement this measure by using employment weights and by calculating HHI_{jt} at the 3-digit NAICS-year level. The results are displayed in the figure below.

Figure 10: National employment concentration has been increasing since the early 1980s.



HHI levels are normalized relative to their initial value in 1976. Source: authors' own calculations from LBD data from 1976 – 2014.

Consistent with Autor et al. (2017), we find that national employment concentration has been rising since the early 1980s. If we look at the whole available period of 1976 – 2014, then it is clear that national concentration has not been rising monotonically. In fact, it was declining from 1976 till 1981 with a particularly sharp drop in 1982 which is consistent with Rinz (2018). While it is tempting to explain this almost continuous drop as measurement error, it is unlikely to be the case with administrative data. Furthermore, Rinz (2018) has argued that it is mainly driven by telecommunications industries and refers to a Department of Justice case in 1982 in which AT&T was required to divest itself of local telephone companies.

Regardless of the rationale behind this drop, it is clear that the time series for national employment concentration does not follow the patterns of our constructed markdown $\mathcal{V}_t$ in the least. Hence, we conclude that caution should be exercised when proxying market power with measures of concentration.

F Concentration in vacancies

We use two sources of data to investigate labor market concentration: employment data from the Longitudinal Business Database (LBD) — as seen in the main body — and vacancy data from Burning Glass Technologies (BGT).

The BGT data is a unique source of micro-data that contains approximately 160 million electronic job postings in the U.S. economy spanning the years 2007 and 2010–2017. These job postings were collected and assembled by BGT, an employment analytics and labor market information company, that examines over 40,000 online job boards and company websites to aggregate the job postings, parse, and deduplicate them into a systematic, machine-readable form, and create labor market analytics products. With the breadth of this coverage, the resulting database purportedly captures the near-universe of jobs posted online, estimated to be near 80 percent of total job ads. Using BGT vacancy data allows us to compute the concentration of job openings, thus zeroing in on concentration in local labor demand and computing an index of concentration that reflects how many employers are active in the hiring process in a local market.

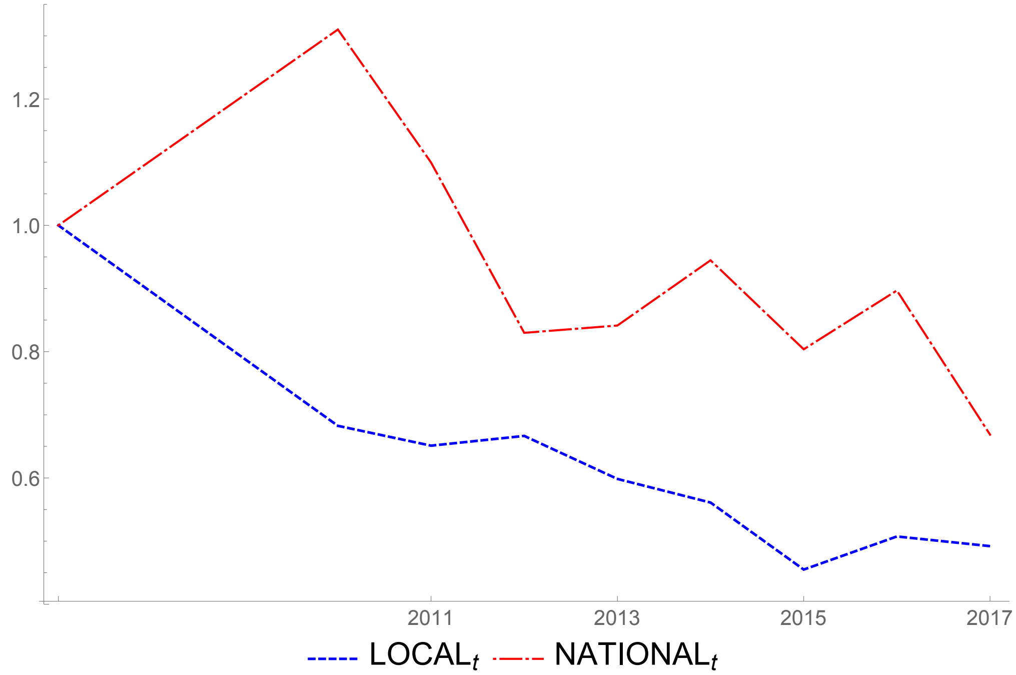
The BGT data has both extensive breadth and detail. Unlike sources of vacancy data that are based on a single job board such as `careerbuilder.com` or `monster.com`, BGT data span multiple job boards and company sites. The data are also considerably richer than sources from the Bureau of Labor Statistics, such as JOLTS (Job Openings and Labor Turnover Survey).[35] In addition to detailed information on occupation, geography, and employer for each vacancy, BGT data contain thousands of specific skills standardized from open text in each job posting. BGT data thus allow for a detailed analysis of vacancy flows within and across occupations, firms, and labor market areas, enabling us to document trends in employers' concentration at a very granular level.

The data, however, is not perfect. Although roughly two-thirds of hiring is replacement hiring, we expect

[35] Although JOLTS asks a nationally representative sample of employers about vacancies they wish to fill in the near term, the data are typically available only at aggregated levels, and do not allow for a detailed taxonomy of local labor markets.

vacancies to be somewhat skewed towards growing areas of the economy (Lazear and Spletzer, 2012; Davis, Faberman and Haltiwanger, 2012). Additionally, the BGT data only covers online vacancies. Even though vacancies for available jobs have increasingly appeared online rather than in traditional sources, it is a valid concern that the types of jobs posted online are not representative of all openings. Hershbein and Kahn (2018) provide a detailed description of the industry-occupation mix of vacancies in BGT relative to JOLTS: although BGT postings are disproportionately concentrated in occupations and industries that require greater skill, the distributions are stable across time, and the aggregate and industry trends in BGT track BLS sources closely.

Figure 11: National and local trends in the concentration of job postings.



HHI levels are normalized relative to their initial value in 2007. Observations from the Great Recession (2008–2009) are not available and are interpolated from 2007 to 2010. Source: BGT (2007, 2010–2017).

In the BGT data, we define a local labor market as an occupation-metro area pair. We define occupations at the 4-digit SOC level, for a total of 108 groups derived from the Bureau of Labor Statistics 2010 SOC system, which aggregates "occupations with similar skills or work activities" (BLS, 2010). While our definition of occupations is considerably less detailed than the job titles available in the BGT data, we believe it offers an appropriate balance between accurately capturing the competitiveness of a market and identifying the demand for different bundles of skills.[36] Nevertheless, our results hold true for other classifications.[37]

[36] Indeed, too fine an occupational classification would mechanically lead to a small number of firms posting jobs in each market. This would bias our estimates of labor market concentration upward. On the other hand, too broad an occupational classification would erase important distinctions between heterogeneous skills used in different occupations. Even though many studies find that broad occupational changes are not uncommon in U.S. labor markets (Huckfeldt, 2017; Macaluso, 2017), especially for laid-off workers, we choose the 4-digit SOC level as a useful compromise.

[37] Examples of SOC 4-digits occupations among Production ones are Food Processing Workers (5130), Assemblers and Fabricators (5120), Textile, Apparel, and Furnishings Workers (5160), and Plant and System Operators (5180).

Metropolitan areas correspond to the 2013 Core-Based Statistical Areas (CBSA) with a population over 50,000. As a result, there are 382 metro areas in our final BGT dataset. In the end, we identify 41,256 local labor markets in the BGT data.

We regard vacancies concentration as the closest measure to the concentration faced by job seekers in a specific (local or national) labor market. We construct local and national concentration measures of vacancies using BGT data. Market-level HHIs are aggregated through their respective vacancy shares.[38] Figure 11 plots the time series of the aggregate local and national concentration of vacancies and shows that local concentration is markedly decreasing over time. Specifically, the local HHI of vacancies drops in the post recession period 2010–2017 by approximately 20 percent. The decrease is even more dramatic if we consider the change between 2007 and 2017 — though it is to be noted that the BGT data is not available during 2008–09. Note that the pattern for the national concentration of vacancies is comparable to its employment counterpart.

[38] Our results are quantitatively unaffected whenever we use employment shares instead.