American Economic Review 2022, 112(7): 2099–2138
https://doi.org/10.1257/aer.20200025

Monopsony in the US Labor Market†

By Chen Yeh, Claudia Macaluso, and Brad Hershbein*

This paper quantifies employer market power in US manufacturing and how it has changed over time. Using administrative data, we estimate plant-level markdowns—the ratio between a plant's marginal revenue product of labor and its wage. We find most manufacturing plants operate in a monopsonistic environment, with an average markdown of 1.53, implying a worker earning only 65 cents on the marginal dollar generated. To investigate long-term trends for the entire sector, we propose a novel, theoretically grounded measure for the aggregate markdown. We find that it decreased between the late 1970s and the early 2000s, but has been sharply increasing since. (*JEL* J24, J31, J38, J42, L13, L60)

Is the US 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. Although a convenient modeling assumption, does this benchmark accurately describe the US 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 US manufacturing plants between 1976 and 2014. Specifically, we show that (i) the US 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 devising appropriate policy responses. Reliable evidence on employer market power is particularly relevant when evaluating

*Yeh: Federal Reserve Bank of Richmond (email: chen.yeh@rich.frb.org). Macaluso: Federal Reserve Bank of Richmond (email: claudia.macaluso@rich.frb.org). Hershbein: W.E. Upjohn Institute for Employment Research (email: hershbein@upjohn.org). Emi Nakamura was the coeditor for this article. 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, Nicholas Trachter, and three anonymous referees, 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 US Census Bureau. We also thank the hospitality of the University of Minnesota, University of Chicago, Federal Reserve Board, and Federal Statistical 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 US 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.
†Go to https://doi.org/10.1257/aer.20200025 to visit the article page for additional materials and author disclosure statements.

policies that directly affect workers' compensation and mobility, such as changes in the minimum wage. Similarly, when assessing regulatory limits on the growth of large firms, it is helpful to consider the extent to which such firms are able to compensate labor below their marginal revenue products. Policymakers have recently considered these policies to mitigate a perceived increase in employers' market power.[1] While this rise in employer market power can be plausibly connected to several labor market trends, measures of employer market power that directly compare the wedge between the marginal revenue product of labor and the wage are not available to inform the current policy debate.[2]

Our paper responds precisely to this gap. We estimate plant-level markdowns for the whole US manufacturing sector and study their relationship with employer size, age, and productivity, and the evolution of aggregate markdowns over time.

Our analysis of labor market monopsony starts with estimating and characterizing the distribution of plant-level markdowns. In our baseline framework, firms internalize a finitely elastic labor supply curve and thus operate in a monopsonistic environment. Without imposing further restrictions on the labor supply curve, we interpret gaps between the output elasticity of labor and labor's revenue share as market power, jointly in output (product markups) and input (labor markdowns). Under the assumption that at least one, other observable input is flexible—that is, free of adjustment costs and monopsony power—we show that markups and markdowns can be identified and estimated separately. The key insight is that the wedge for the flexible input reflects only product markups, and so the *ratio* of the labor wedge and the wedge for the flexible input permits identification of both markups and labor markdowns. To implement this insight empirically, we adapt the production function approach from the industrial organization (IO) literature.

This approach has several advantages. First, we can remain agnostic about the sources of employer market power. In fact, we show that our approach is consistent with a broad range of monopsony models. Second, although we do need to impose a functional form for a firm's production, we can be highly flexible by specifying a translog function—a second-order approximation to *any* arbitrary, differentiable production function. A third benefit is that the production function approach remains valid regardless of the assumptions made on other inputs besides labor and materials (and thus can accommodate capital adjustment costs). Finally, the approach readily permits several extensions and modifications, including heterogeneous labor within plants, labor adjustment costs, ex ante specified returns to scale, and alternative measures of labor compensation, such as inclusion of benefits.

Estimating such production functions and markdowns with comprehensive administrative data for the US manufacturing sector (using the Census of Manufactures and Annual Surveys of Manufactures), we find that that labor markets in US manufacturing are far from perfectly competitive. The average plant's marginal revenue

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

[2] Several complementary measures have been proposed, including those related to labor market concentration (Azar, Marinescu, and Steinbaum 2020; Azar et al. 2020; Benmelech, Bergman, and Kim 2020; Rinz 2020; Schubert, Stansbury, and Taska 2021), as well as fully structural approaches (Posner, Weyl, and Naidu 2018; Azar, Berry, and Marinescu 2019; Jarosch, Nimczik, and Sorkin 2021; Berger, Herkenhoff, and Mongey 2022). The measure we develop is based on the production function approach and is unique in that it quantifies, with minimal assumptions, plant-level wedges between the marginal revenue product and the wage.

product of labor is 53 percent higher than its wage, implying that a worker employed there receives about 65 cents on the marginal dollar. Furthermore, we document a substantial amount of dispersion across plants even within three-digit NAICS industries, with an average within-industry interquartile range of 61.6 percent. Investigating the sources of heterogeneity in markdowns, we find a robust positive association between markdowns and size, whether measured as an establishment's (or firm's) relative share of employment or in terms of industrial and geographical scope. This result supports the hypothesis that employer size matters when assessing the welfare implications of labor market power. On the other hand, we find that plant-level dispersion in markups and in productivity account for little of the heterogeneity in markdowns. We conclude that the typical US manufacturing plant operates in a monopsonistic environment, and a significant degree of variation persists within narrowly defined industries.

We next use our estimates of micro-level markdowns to describe trends in macro-level markdowns since 1977. This is not straightforward, as there is no uncontested framework that delivers a clear aggregation rule for markdowns. We propose a novel "aggregate markdown" measure that satisfies two requirements. First, aggregate markdowns and markups reflect *aggregate* wedges, the 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 for aggregation. Second, aggregate markdowns need to account for the local nature of labor markets, 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. This measure of the aggregate markdown displays a U-shaped evolution over time, decreasing between 1977 and 2002, and sharply increasing afterwards. We thus find support for the hypothesis that monopsony in the US manufacturing labor market has increased since the early 2000s.

Finally, we relate our aggregate markdown estimates to measures of labor market concentration that have been commonly used in recent studies, as described below, on account of their simplicity of construction. Despite our finding that *plant-level* markdowns increase with employment size, we find little correlation between concentration and markdowns at the aggregate, *market* level, a consequence of accounting for heterogeneity in output elasticities across firms within a market and our aggregation rule. Furthermore, although aggregate local concentration and our aggregate markdowns show qualitatively similar declines in the late twentieth century, the concentration measure does not show the sharp reversal of markdowns since the early 2000s. We conclude that cross-sectional and time variation in local employment concentration do not necessarily reflect variation in employer market power as measured by markdowns—at least within manufacturing.

Contribution to the Literature.—Our paper contributes to a recently reinvigorated research agenda on the prevalence and evolution of labor market monopsony in the US economy. This interest in the exercise of market power by firms, and especially large firms, has been motivated heavily by the secular decline in labor's share of income (Elsby, Hobijn, and Şahin 2013; Karabarbounis and Neiman 2013), which has in turn been linked to changes in industry-level sales concentration, with "superstar" firms potentially having higher product markups and lower labor shares (Autor et al. 2020).[3]

Our contribution is twofold. First, we use comprehensive administrative data for US manufacturers and provide direct estimates of the wedge between an employer's marginal revenue product of labor and its wage. In so doing, we document substantial dispersion in plant-level markdowns and document how this heterogeneity varies with employer characteristics, such as size, age, productivity, and geographic scope. Second, we develop a new, theory-grounded way to characterize aggregate markdowns and document their evolution over the past four decades.

Our estimation procedure to obtain markdowns relies on the "production approach" (De Loecker 2011), which combines insights from Hall (1988) with production function estimation techniques from the IO literature (Olley and Pakes 1996; Levinsohn and Petrin 2003; De Loecker and Warzynski 2012; Ackerberg, Caves, and Frazer 2015). 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 possibly overstate the extent of that source's market power. Exceptions to this practice, however, include Dobbelaere and Mairesse (2013) and Morlacco (2020), who also exploit the flexibility of material inputs to study monopsony in, respectively, non-US labor markets and the market for foreign intermediate inputs. Brooks et al. (2021b) also use techniques analogous to those in this paper to estimate markdowns in China and India, and conclude that "in the context of developing economies, markdowns substantially lower the labor share."[4]

A related literature has documented a contemporaneous increase in markups, arguing the latter could be a unifying explanation behind many observed secular trends in the US economy, including the decrease in the labor share (De Loecker, Eeckhout, and Unger 2020; Eggertsson, Robbins, and Wold 2021). Our paper 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 between markdowns and size.

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 (1999); Manning (2003); Mortensen (2003); Manning (2011); and Webber (2015). It also encompasses 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 (2022)—and environments based on monopsonistic competition, as in Bhaskar and To (1999); Staiger, Spetz, and Phibbs (2010); and Berger, Herkenhoff, and Mongey (2022). Our paper contributes to this literature by proposing a strategy to estimate markdowns that,

[3] With the exception of the lowest-productivity establishments, which we discuss further below, we find a positive relationship between plant-level productivity and markdowns. This finding is consistent with the thesis in Autor et al. (2020) as long as labor shares fall faster than product markups rise.

[4] In a companion paper, Brooks et al. (2021a) further show that highway construction in India offset markdowns and increased the labor share among nearby firms.

while compatible with many of the frameworks studied previously, is not tightly linked to a specific micro-foundation but instead is quite general.[5]

Finally, our paper relates to the burgeoning literature on labor market concentration, as we compare concentration indices to markdowns. Interest in concentration indices stems from their ease and breadth of use in both academic research and the practice of antitrust in the US economy. These have been calculated at the national (Autor et al. 2020) and local levels (Rossi-Hansberg, Sarte, and Trachter 2020), and show diverging long-run trends.[6] Recent work by Azar, Marinescu, and Steinbaum (2020) and Azar et al. (2020) show the negative association between concentration and wages using vacancy data from online sources and argue for extending antitrust considerations to mergers that affect labor market concentration. 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 conversation by documenting that the correlation between markdowns and employment concentration is quite modest, both cross-sectionally (across local labor markets) and in the aggregate over time. We view this result as highlighting the challenges posed by aggregation when comparing micro-founded measures of employer market power, such as markdowns, to reduced-form indices, such as employment concentration.

Overview of This Paper.—Section I lays out our estimation procedure and describes the data. Section II illustrates our markdown estimates and discusses heterogeneity. Section III 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. In Section IV, we discuss the robustness of our baseline results. Section V summarizes the evidence and concludes. We provide several additional results, derivations, and robustness tests in the Appendix and online Appendix.

I. Markdown Estimation

Our analysis of monopsony in the US labor market is based on markdowns, the percentage gap between a firm'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. When markdowns are larger than unity, however, the employer compensates workers less than dollar-for-dollar for every unit of revenue generated at the margin.

[5] 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. Our implied average elasticity estimate of 1.88 is only slightly above the median elasticity estimate from more than 800 research papers covered in the meta-study by Sokolova and Sorensen (2021).

[6] A recent paper by Benmelech, Bergman, and Kim (2020) also computes employment concentration and relates it to average wages in US manufacturing. Lipsius (2018) and Rinz (2020) 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 US economy.

In this section, we describe our basic framework. We begin by using the optimality conditions from a firm's profit maximization problem to show a one-to-one relationship between markdowns and firm-level labor supply elasticity. We then use the firm's dual problem (through cost minimization) to derive an estimation strategy for markdowns 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 US manufacturing sector. Our approach simultaneously allows for positive product markups.

A. *Obtaining Markdowns through Duality*

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). We refer to this percentage gap as a firm's markdown. In the following, we will show that a firm's markdown has a one-to-one relationship with its (perceived) labor supply elasticity.[7] To see this, consider a firm's profit maximization problem,

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

where $R(\ell) \equiv \text{rev}\big(\ell; \mathbf{X}^*_{-\ell}(\ell)\big)$ 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

$$(1) \qquad R'(\ell^*) = \left[\frac{w'(\ell^*)\,\ell^*}{w(\ell^*)} + 1\right] w(\ell^*)$$
$$= \left[\varepsilon_S^{-1} + 1\right] \cdot w(\ell^*),$$

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

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

so that the markdown, ν is expressed as the ratio of the MRPL to the wage, or the inverse labor supply elasticity plus one.[8] In this conceptual framework, we do not take a specific stance on the sources of monopsony power; the only requirement is

[7] This parallels the intuition behind the Lerner index formula, which relates residual demand elasticities with price-cost markups.

[8] Some studies define the markdown as the inverse of our measure since it reflects the extent to which wages are marked down. Under this convention, markdowns below unity reflect labor market power, whereas in our measure, markdowns *above* unity reflect labor monopsony.

finite firm-specific labor supply elasticities. In online Appendix O.8, we show that our setup is quite general and nests a variety of monopsony frameworks, including wage-posting models (e.g., Burdett and Mortensen 1998), additive random utility models (e.g., Card et al. 2018; Chan, Kroft, and Mourifie 2019; and Lamadon, Mogstad, and Setzler 2022), and monopsonistic competition models (e.g., Bhaskar and To 1999; Staiger, Spetz, and Phibbs 2010; and Berger, Herkenhoff, and Mongey 2022).

Cost Minimization.—A complication, however, is that estimating a firm's perceived elasticity of labor supply in a general setting is challenging, in part because of the potential for firm market power over both inputs (monopsony) and output (monopoly). In this section, we propose a "production approach" to retrieve markdowns for US manufacturers in a general setting, building on insights from Hall (1988); De Loecker (2011); and De Loecker and Warzynski (2012). The key insight is that wedges between output elasticities and revenue shares can reflect market power in both input and output markets. Intuitively, the output elasticity of labor captures the gain from 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.

The production approach starts with a firm's optimal input choices. Suppose there are $K > 1$ inputs, denoted by $\mathbf{X}_{it} = \left(X_{it}^1, \ldots, X_{it}^K\right)'$. These inputs have pricing schedules $\left\{V_{it}^k\right\}_{k=1}^K$, and adjustment costs for some input k are captured by the function $\Phi_t^k\left(X_{it}^k, X_{it-1}^k\right)$. Also, denote a firm i's productivity level at time t by ω_{it}. Then, to derive markdowns, we adopt the following set of assumptions.

ASSUMPTION I: *A firm engages in cost minimization.*

There exists at least one input k' that satisfies the following:

ASSUMPTION II: *There are no adjustment costs for input* k', *i.e.*, $\Phi_t^{k'}(\cdot, \cdot) = 0$.

ASSUMPTION III: *Input* k' *is not subject to monopsony forces*, *i.e.*, $V_{it}^{k'}\left(X_{it}^{k'}\right) = V_{it}^{k'}$.

ASSUMPTION IV: *Input* k' *is chosen statically.*

ASSUMPTION V: *Production* $F(\cdot;\, \omega_{it})$ *is twice continuously differentiable in* $X_{it}^{k'}$ *and it satisfies*

$$\lim_{X_{it}^{k'} \to 0} \frac{\partial F\left(\mathbf{X}_{it}; \omega_{it}\right)}{\partial X_{it}^{k'}} = +\infty \quad \text{and} \quad \lim_{X_{it}^{k'} \to +\infty} \frac{\partial F\left(\mathbf{X}_{it}; \omega_{it}\right)}{\partial X_{it}^{k'}} = 0.$$

for any $\omega_{it} \in \mathbb{R}_+$. *Furthermore*, *the demand schedule* $P_{it}(\cdot)$ *is continuously differentiable and strictly decreasing*.

ASSUMPTION VI: *Input* k' *is used for the production of output only.*

Any input k' that satisfies Assumptions II–VI simultaneously, will be referred to as a *flexible* input. Assumptions I and V are regularity assumptions, and ensure that (a subset of) inputs can be characterized through their first order conditions alone. Assumption VI is relatively weak and requires the flexible input to be used solely for production purposes.[9] Assumption IV implies that an input k' cannot directly affect a firm's future outcomes, ruling out certain dynamic narratives.[10]

Hence, the challenge becomes to find an input that simultaneously satisfies Assumptions II and III: the existence of a static input k' free of adjustment costs and for which firms are price-takers. If such an input k' exists, we can establish the following result.

PROPOSITION 1: *Let Assumption I hold. If Assumptions II–VI hold for some input k' other than labor, we can characterize a firm's product markup with the gap of its flexible input k': $\mu_{it} = \frac{\theta_{it}^{k'}}{\alpha_{it}^{k'}}$ where $\theta_{it}^{k'}$ and $\alpha_{it}^{k'}$ denote a firm's output elasticity of input k' and its share of revenue, respectively. If Assumptions II and IV–VI also apply to labor ℓ, and firm i faces a differentiable, finitely elastic wage schedule, then its markdown ν_{it} satisfies*

$$\nu_{it} = \frac{\theta_{it}^{\ell}}{\alpha_{it}^{\ell}} \cdot \mu_{it}^{-1}, \tag{3}$$

where θ_{it}^{ℓ} and α_{it}^{ℓ} denote a firm's output elasticity of labor and its labor share of revenue, respectively.

PROOF:

See Appendix A.

Proposition 1 implies that the ratio between the output elasticity of labor and labor's share of revenue equals the product of the markdown and the markup. In the remainder of this paper, we follow the IO literature and assume that conditions II–VI hold for material inputs (therefore referred to as "the flexible input" and indexed by M). The availability of a flexible input is the key factor that allows us to distinguish between markdowns and markups, isolating our measure for labor market power from market power for outputs.

Thus, a key question for our identification is whether materials lack adjustment costs and monopsony power in our context. Several pieces of evidence suggest these are reasonable assumptions. First, the US Census Bureau's definitions of material inputs includes largely generic, primary goods, as well as contract services, that tend to be traded on open, often global markets.[11] Second, Atalay (2014) does not find

[9] This rules out, for example, inputs designed purely to increase the demand for output, such as marketing. Because our data are at the establishment level and allow us to separate production from nonproduction labor, this assumption is relatively innocuous, and we discuss it in detail in Section IV.

[10] Our setup allows for (dynamic) capital adjustment costs as long as there is a flexible input other than capital. We do rule out, however, mechanisms in which current output (which depends on current inputs) affects a firm's future demand, such as by affecting the customer base (see, for example, Foster, Haltiwanger, and Syverson 2016).

[11] See Table 11 in online Appendix O.6 for a complete list.

that prices for material inputs vary with quantities (as required by Assumption III).[12] Third, even if material inputs were subject to monopsony forces, our estimates would reflect markdowns for labor *relative* to markdowns for material inputs, implying our estimates would be lower bounds for labor market power.[13]

For labor markdown estimation, Proposition 1 also requires Assumptions II and IV–VI apply to *labor*. For Assumptions II and IV to hold, we need to rule out labor adjustment costs, including non-spot-market contracts.[14] We address the sensitivity of our results to possible labor adjustment costs in online Appendix O.4 and find that these matter relatively little, for both convex and nonconvex adjustment costs.

Finally, our setup also excludes labor being used for any purpose other than the production of output (Assumption VI). This could occur, for example, when labor is used for marketing or hiring. This assumption may seem strong, but it is less restrictive in our context of US manufacturing plants, where most workers are indeed production workers. Furthermore, we show in Section IV that our results are not affected when we explicitly distinguish production from nonproduction labor.

B. *Production Function Estimation*

A distinct advantage of the production approach is its generality. 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 from the fully developed structural models of Card et al. (2018); Chan, Kroft, and Mourifie (2019); Lamadon, Mogstad, and Setzler (2022); and Berger, Herkenhoff, and Mongey (2022). Furthermore, we need make no additional assumptions for inputs besides materials and labor. Our approach is valid as long as firms are subject to some finitely elastic labor supply curve and material inputs are flexible. Finally, we can explicitly distinguish between market power in output and labor markets. Our result in Proposition 1 implies that observing output elasticities and revenue shares is sufficient for constructing markdowns. Revenue shares are directly observable in administrative data on US manufacturing plants, but output elasticities need to be estimated.

To do so, we estimate production functions via "proxy variable" methods (Olley and Pakes 1996; Levinsohn and Petrin 2003; De Loecker and Warzynski 2012; Ackerberg, Caves, and Frazer 2015). We adopt standard assumptions from the proxy variable literature, particularly Ackerberg, Caves, and Frazer (2015), which we discuss below.

[12] Atalay (2014) also finds that some US manufacturing plants pay "low materials prices because their suppliers are exceptionally productive." Hence, the variation in material input prices across plants can be partially explained by variation in the marginal costs of suppliers, rather than plants exploiting their monopsonistic power.

[13] We discuss these issues in more detail in Section IV.

[14] Adjustment costs in labor would occur, for instance, in the presence of hiring and firing costs, such as a corporate tax schedule that varies with firm size, or legal requirements that limit employment at will. However, these provisions are not especially binding in the US labor market.

ASSUMPTION 1: *A firm i's information set at time t,* $\mathcal{F}_{it}$*, is generated by* $\{\omega_{i\tau}\}_{\tau=0}^{t}$*. The transitory shock* ε_{it} *is not observed by the firm and satisfies* $\mathbb{E}[\varepsilon_{it} \mid \mathcal{F}_{it}] \equiv \mathbb{E}_t(\varepsilon_{it}) = 0$.

ASSUMPTION 2: *A firm i's state variables at time t are given by the pair* (k_{it}, ω_{it})*. Furthermore, its stock of capital accumulates as a function of lagged capital* k_{it-1} *and investment* ι_{it-1}:

$$k_{it} = \kappa(k_{it-1}, \iota_{it-1}). \tag{4}$$

ASSUMPTION 3: *The technology parameters* $\boldsymbol{\beta}$ *are constant across time and common within an industry group. Productivity evolves according to a first-order Markov process:*

$$p(\omega_{it+1} \mid \mathcal{F}_{it}) = p(\omega_{it+1} \mid \omega_{it}), \tag{5}$$

where the distribution $p(\cdot \mid \mathcal{F}_{it})$ *is known to firms and is stochastically increasing in* ω_{it}.

ASSUMPTION 4: *For at least one flexible input, the only unobservable factor (from the econometrician's point of view) in a firm's input demand function is productivity* ω_{it}*, which is a scalar and Hicks-neutral.*

ASSUMPTION 5: *For any flexible input satisfying Assumption 4, a firm's input demand function is strictly monotone in* ω_{it}.

Assumptions 1 and 2 are standard in the literature and encompass a rich set of frameworks of firm behavior. Because the econometrician does not observe firm-level productivity, a least squares regression of output on inputs would lead to biased estimates (the "transmission bias" in Griliches and Mairesse 1998). To address this problem, Assumption 3 places some general structure on idiosyncratic productivity by having it follow a first-order Markov process.[15]

Assumption 4, also known as the "scalar unobservable" assumption, requires that idiosyncratic productivity be the only input demand factor unobserved by the econometrician. Bond et al. (2021) argue this assumption is not consistent within the context of market power, and other estimators that do not rely on it should be used instead (e.g., Blundell and Bond 2000). However, in online Appendix O.6, we run a set of Monte Carlo simulations and demonstrate that our empirical approach, relying on the proxy variable estimator of De Loecker and Warzynski (2012), still produces estimates with less bias than do estimators that do not rely on the scalar unobservable assumption. Consequently, we do not view Assumption 4 as restrictive in practice.

[15] Although not without loss of generality, this structure is the most general in the production function estimation literature (Ackerberg 2020) and allows for considerably greater flexibility than other common persistent stochastic processes, such as an AR(1).

Finally, we require that flexible input demand functions are invertible in productivity. This assumption allows us to obtain a nonparametric estimate of output without observing productivity. While implicitly ruling out some production functions, this assumption still allows for a general specification such as the translog production function that we adopt in our baseline estimates.[16] Its flexibility rests on its interpretability as a second-order approximation to *any* arbitrary, differentiable 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.[17]

In the following, we briefly describe the mechanics of the proxy variable methodology and how we obtain output elasticities. We refer the reader interested in a more detailed treatment of our estimation procedure to online Appendix O.3. In Section IV, we further discuss possible challenges to the proxy variable methodology and how they relate to our results.

We denote y_{it} as log output and $\mathbf{x}_{it}$ as the vector of log inputs. This vector of inputs contains the first-, cross-, and second-order terms of the vector $\tilde{\mathbf{x}}_{it} = (k_{it}, \ell_{it}, m_{it}, e_{it})'$ consisting of capital, labor, materials, and energy. Because of the unobserved productivity parameter, we require instruments for the input vector to recover consistent production parameters. Let $\mathbf{z}_{it}$ be the vector instrumenting for the set of endogenous inputs $\mathbf{x}_{it}$. Following De Loecker and Warzynski (2012), we construct $\mathbf{z}_{it}$ by taking the lag of each input in $\mathbf{x}_{it}$ with the exception of capital. Last, let $f(\mathbf{x}_{it};\boldsymbol{\beta})$ denote the log transformation of the production function. In the end, our goal is to estimate production function parameters $\boldsymbol{\beta}$ in the following setting:

$$y_{it} = f(\mathbf{x}_{it};\boldsymbol{\beta}) + \omega_{it} + \varepsilon_{it}, \tag{6}$$

where ε_{it} reflects measurement error.[18] The loglinearity of output in productivity comes from the second part of Assumption 4. Given Assumptions 1–5, we estimate production function parameters $\boldsymbol{\beta} \in \mathbb{R}^Z$ for each industry-specific production function in a three-step process:

1. Run a third-order polynomial regression of y_{it} on the inputs $\tilde{\mathbf{x}}_{it}$ and interactions, as well as a set of controls.[19] Obtain nonparametric estimates of log output φ_{it} free of measurement error.

[16] This is important because the production approach does require a functional form on a firm's production function. For our estimates to have some external validity, it is therefore desirable to adopt a production structure that is as general as possible. We believe this is achieved through a translog specification.

[17] Assumption 2 allows production parameters to vary across detailed industry groups (i.e., three-digit NAICS) but imposes that they are constant over time. However, in our context with translog production 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. 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.

[18] Other interpretations for ε_{it} are possible (Bond et al. 2021). We follow De Loecker and Warzynski (2012) and interpret ε_{it} as measurement error. Its exact interpretation is not important for our results as long as ε_{it} is unobserved by the firm and the econometrician.

[19] Our baseline estimates include a set of year fixed effects, but our results do not change by much when other controls, such as size and age, are included.

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

3. Obtain estimates $\hat{\boldsymbol{\beta}}$ of the production function parameters $\boldsymbol{\beta}$ through the GMM system induced by the moment conditions $\mathbb{E}(\xi_{it}(\tilde{\boldsymbol{\beta}}) \cdot \mathbf{z}_{it}) = \mathbf{0}_{Z\times 1}$.

Once estimates of $\boldsymbol{\beta}$ 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 $\tilde{\mathbf{x}}_{it}$, with coefficients that depend on $\boldsymbol{\beta}$. A complete description on the construction of output elasticities under translog production is found in online Appendix O.3.

A crucial part of the proxy variable methodology is to obtain transitory shocks to firm-level productivity. As a result, we need to separately identify productivity ω_{it} and measurement error ε_{it}. Under Assumptions 4 and 5, productivity can be written as a function of observables only. This allows us to identify ε_{it} in the first step.[20] Our translog structure then allows us to obtain firm-level productivity in step two. Finally, we are able to identify transitory shocks ξ_{it} to productivity through the Markov property in Assumption 3. These shocks are the key behind our moment conditions: current inputs are orthogonal to future shocks in productivity through Assumption 2.

C. *Intuition behind Identification*

The econometric literature on production function estimation has not provided formal arguments on whether proxy variable estimators produce consistent estimates. However, informal arguments for identification can be given through the logic of an IV estimator. As demonstrated in step 3 of our estimation procedure, we construct our moment conditions through the instrument vector $\mathbf{z}_{it}$. Therefore, we are "instrumenting" the endogenous input vector $\mathbf{x}_{it}$ with $\mathbf{z}_{it}$. To understand why the above system can retrieve valid estimates of $\boldsymbol{\beta}$, we can verify a set of exogeneity and rank conditions.

Exogeneity implies that $\mathbf{z}_{it}$ is orthogonal to the innovations to productivity in period t. By Assumption 2, firms choose inputs k_{it}, ℓ_{it-1}, m_{it-1}, and e_{it-1} without knowing what the productivity shock ξ_{it} will be. As a result, exogeneity holds through our timing assumptions. However, how do we ensure that the "instruments" are valid, that those moment conditions associated with ℓ_{it-1}, m_{it-1} and e_{it-1} are actually informative for the production function coefficients on inputs ℓ_{it}, m_{it}, and e_{it}? For this to follow, we need factor input demand to evolve relatively smoothly, ruling out, for example, production functions that reflect perfect substitutes or perfect complements. We also need price schedules to similarly evolve smoothly and be

[20] By Assumptions 1, 2, and 4, we can write $m_{it} = m_t(\omega_{it}; k_{it})$. Whenever Assumption 5 also holds, m_t is invertible in productivity and there exists some function $\omega_{it} = h_t(m_{it}; k_{it})$. Thus, we have $y_{it} = f(\mathbf{x}_{it}; \boldsymbol{\beta}) + h_t(m_{it}; k_{it}) + \varepsilon_{it} \equiv \phi(\mathbf{x}_{it}; \gamma) + \varepsilon_{it}$. Hence, we can obtain estimates for output net of measurement error by running a nonparametric regression (e.g., a high-order polynomial) in only observables.

somewhat—but not perfectly—persistent over time. We believe these are plausible requirements; Atalay (2014), for instance, finds empirical support for the partial persistence of materials prices.

For consistent estimates, we further require reasonably long panel data (Pesaran and Smith 1995; Gandhi, Navarro, and Rivers, 2020; Bond et al. 2021). In particular, Gandhi, Navarro, and Rivers (2020) formalize that time-series variation in the prices for material inputs—our flexible input—is critical for identification, as it is the *only* residual source of variation that can identify β under the proxy variable methodology.[21] As our data, described below, span almost 40 years, we believe there is sufficient variation in materials prices, whether in aggregate or industry-specific, to identify production function parameters.[22]

D. *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 US Census Bureau. The Census of Manufactures is a quinquennial survey that covers the universe of manufacturing establishments in years ending in "2" and "7." Crucially, the CM contains establishment-level data on revenues and inputs, the two necessary ingredients for production function estimation. We construct our measures of output (revenues) and inputs (capital, labor, materials, and energy) using deflators from the NBER-CES Manufacturing Database, following the standard procedures described in Syverson (2004a) and Kehrig (2015).

To construct markdowns for non-census years, we use the Annual Survey of Manufactures (ASM). The ASM contains a representative, rotating sample of manufacturing plants. While large plants are sampled with near certainty, small plants are sampled less frequently based on their size.[23] We use survey-provided sampling weights to ensure that our estimates are representative of the whole manufacturing sector. Our main results are thus based on a nonbalanced panel for manufacturing plants in years 1976–2014. To avoid artificial spikes in census years, we keep only those plants that are in the rotating sample of the ASM in these years.

[21] As Flynn, Gandhi, and Traina (2019) point out, this input price variation should be orthogonal to productivity and output prices. As a result, some forms of unobserved heterogeneity in inputs can be problematic. For example, it cannot reflect differences in the quality of purchased inputs. While it is difficult to verify these assumptions explicitly in the absence of input quantity data, we have verified that the overwhelming majority of the variation in material input deflators from the NBER-CES Manufacturing Database comes from the time series.

[22] We should note that our measures of output and inputs are based on deflated expenditures. While we show that markdowns can be obtained even if only (deflated) revenue elasticities can be estimated (see Section IV), the absence of input prices does technically violate the scalar unobservable assumption (see Hu, Huang, and Sasaki 2020; Bond et al. 2021). However, this issue does not seem to be a major concern in practice, as we summarize in Section IVG. Using Monte Carlo methods, we show more fully in online Appendix O.6 that our preferred proxy variable estimator appears more robust than other estimators that do not rely on the scalar unobservable assumption.

[23] Plant size is determined in terms of revenues and/or employment.

Table 1—Summary Statistics of Plant-Level Markdowns

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
Motor vehicles	1.368	1.422	0.376	0.432
Printing and publishing	1.345	1.495	0.454	0.632
Electrical machinery	1.317	1.416	0.519	0.513
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
Nonmetallic 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	$1.393 \cdot 10^6$			

Notes: 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 US Bureau of Economic Analysis, which approximately follows a three-digit NAICS specification. The distributional statistics are calculated using sampling weights provided in the data.

Source: Authors' calculations from ASM/CM data in 1976–2014.

II. Markdowns in US Manufacturing

A. *Cross-Sectional Distribution*

We present results of our estimation procedure in Table 1. These paint a clear picture: markdowns are sizable and considerably larger than unity. The average establishment throughout the period 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. Alternatively, taking the reciprocal, a markdown of 1.53 implies that a worker receives about 65 cents on the marginal dollar generated.

Furthermore, we find that labor market power is widespread across manufacturing plants. One-half charge a markdown of at least 1.364 (73 cents on the dollar), and the interquartile range in markdowns exceeds 0.6. Although these markdown estimates may seem large, they are largely in line with implied estimates from previous studies (see Manning 2003; Webber 2015; Sokolova and Sorensen 2021). When we compare our markdown estimates with the meta-analysis on labor supply elasticities by Sokolova and Sorensen (2021), our estimates fall below the median 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, there is considerable variation in markdowns across plants *within* the same industry. The average *within-industry* interquartile range (standard deviation) of markdowns is 61.6 (60.4) percent. This suggests that heterogeneity in markdowns likely relates 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.[24]

Recent studies have emphasized that the welfare cost of market power distortions can be considerable (Baqaee and Farhi 2020; Edmond, Midrigan, and Xu 2021; Berger, Herkenhoff, and Mongey 2022), and so we next turn to understanding the determinants of markdown variation.

B. *Heterogeneity in Markdowns*

Variance Decomposition.—To investigate markdown heterogeneity, we first decompose markdowns into their components according to equation (3). Micro-level markdowns are additively separable (in natural logs) according to

$$\ln(\nu) = \ln(\theta_\ell) - \ln(\alpha_\ell) - \ln(\mu). \tag{7}$$

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:

$$\begin{aligned} V\big(\ln(\nu)\big) = {} & V\big(\ln(\theta_\ell)\big) + V\big(\ln(\alpha_\ell)\big) + V\big(\ln(\mu)\big) \\ & -2 \cdot \Big[\mathrm{cov}\big(\ln(\theta_\ell),\ln(\alpha_\ell)\big) - \mathrm{cov}\big(\ln(\alpha_\ell),\ln(\mu)\big) \\ & + \mathrm{cov}\big(\ln(\theta_\ell),\ln(\mu)\big)\Big]. \end{aligned} \tag{8}$$

In Table 2, 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, as well as their covariance. Variations in markups, on the other hand, play a quantitatively small role for markdown variation.[25] Our results consequently imply that the main determinants of markdown variation are different from those that drive variation in markups.

Size, Age, and Productivity.—We proceed to investigate the source(s) of markdown variation by focusing on idiosyncratic factors, especially those likely to be related to labor supply elasticities and labor shares. 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 (Edmond, Midrigan, and Xu 2021; Berger, Herkenhoff, and Mongey 2022), and it

[24] This pattern accords with the dispersion in revenue-based total factor productivity documented by Syverson (2004b). In online Appendix O.1, we explore whether *industr*y-level characteristics (e.g., unionization) can explain some of the observed heterogeneity in markdowns. Qualitatively, we find a slight negative relationship between markdowns and unionization at the industry-state level.

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

Table 2—Variance Decomposition of Plant-Level Markdowns

		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	−3.307
	θ_ℓ, μ	−0.00601	0.0709
	α_ℓ, μ	−0.00271	−0.0320

Notes: Variance decomposition of markdowns as based on equation (8). Relative contributions are calculated as a fraction of the markdown's total variance. Relative contributions of covariance terms have been multiplied by a factor two as consistent with equation (8).

Source: Authors' calculations from ASM/CM data in 1976–2014.

is thus 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. We therefore run a set of nonparametric regressions to flexibly capture the heterogeneity of markdowns by size and age. These regressions are of the following form:

$$\nu_{it} = \beta_0 + \sum_{d=1}^{S} \beta_d^{\text{size}} \cdot \mathbf{1}_{s_{it} \in S_d} + \sum_{d=1}^{A} \beta_d^{\text{age}} \cdot \mathbf{1}_{\text{age}_{it} \in A_d} + \mathbf{X}_{it}' \gamma + \varepsilon_{it}, \tag{9}$$

where $\mathbf{X}_{it}$ contains a full set of industry, state, and year fixed effects. We create size dummies over a grid of $S = 10$ equally spaced bins of a plant's employment share of its local labor market (NAICS3-county cell).[26] We categorize age into 8 groups.[27] The results, depicted in Figure 1, display a clear picture: markdowns are monotonically increasing in size. Conditional on plant age, industry, and other covariates, markdowns for plants with the highest shares of employment are, on average, roughly 20 percent higher than for the smallest plants.

The results for age are somewhat similar but less clear-cut. Without size controls, there is a statistically significant positive age gradient in markdowns. However, as shown in Figure 2, this relationship is attenuated once size controls are included, and we cannot reject that average markdowns are similar across the plant age distribution. Consequently, the relationship between markdowns and plant age is not especially robust.

We also investigate the relationship between markdowns and plant-level productivity. Previous studies have identified a positive association between wages and

[26] Following Haltiwanger, Jarmin, and Miranda (2013), we apply employment weights. However, our results are little affected if we do not use employment weights.

[27] These age groups include: 0–2 years, 3–4, 5–6, 7–8, 9–10, 11–12, 13–15, and 16+ years. To minimize reporting bias in size and age, we take a plant's employment share and age from the Longitudinal Business Database (LBD), which contains the universe of employers, and merge them to ASM/CM at the establishment-year level.

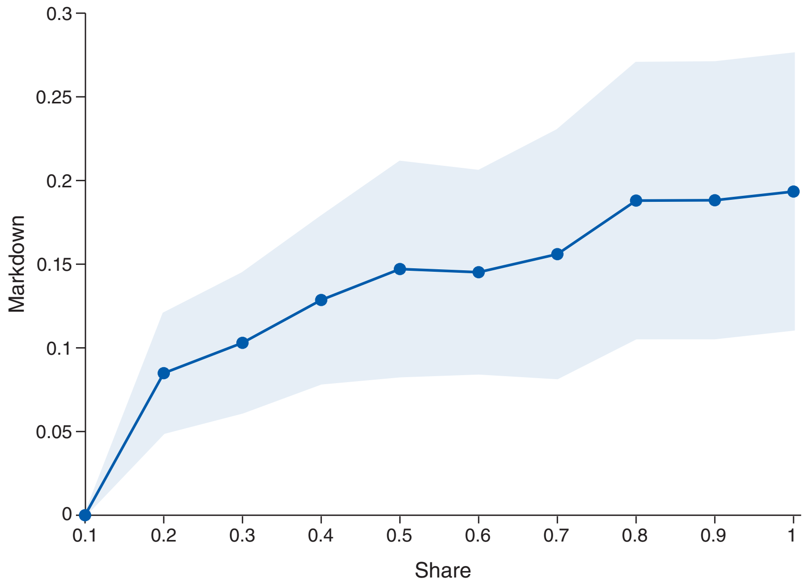

Figure 1. Markdown-Size Relationship

Notes: The figure shows point estimates and 95-percent confidence intervals of plant-specific markdowns on size (as measured by employment share) indicators, controlling for indicators for plant age and industry, as well as state and year fixed effects. The omitted group is the smallest size indicator, so coefficients reflect deviations relative to this baseline. The indicator labeled "0.1" is equal to unity for those plants with employment shares $s \in (0, 0.1]$. Other indicators are defined similarly. Standard errors are clustered at the industry level.

Source: Authors' own calculations from ASM/CM data in 1976–2014.

profits (or sales) per worker. Christofides and Oswald (1992), for example, 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. Similarly, strategies popularized by Abowd, Kramarz, and Margolis (1999) have found positive sorting and correlation between workers' bargaining power and firms' profitability measures to partially explain the positive relationship between wages and productivity. More recently, Card, Devicienti, and Maida (2014) estimate an elasticity of wages to (economic) rents of approximately 4 percent, and Seegmiller (2021), using a dynamic wage posting model, finds that public firms higher in the labor productivity distribution have greater markdowns. In a related vein, 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. Since we do not observe quantities, we proxy physical productivity (TFPQ) by revenue productivity (TFPR).[28]

Unlike that for size and age, we find the relationship between a plant's markdown and productivity is not monotonic. As shown in Figure 3, the data suggest more of a U-shaped association between markdowns and productivity. Markdowns are

[28] Although being able to observe TFPQ would be ideal, 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.

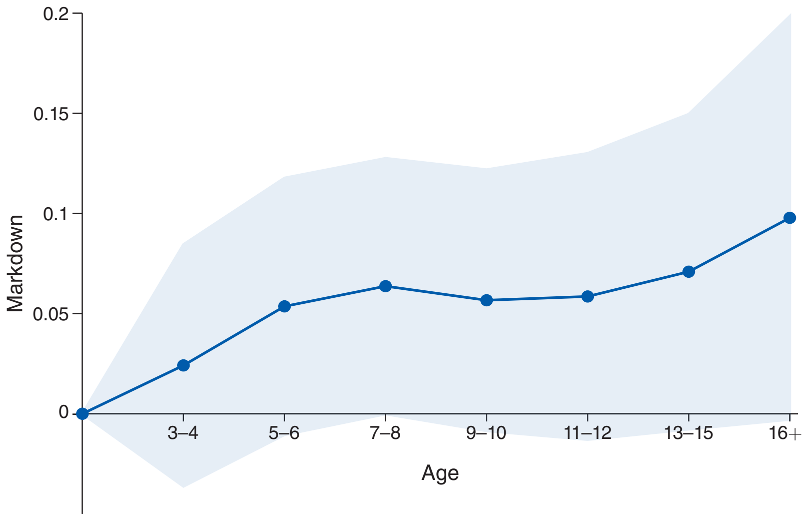

Figure 2. Markdown-Age Relationship

Notes: The figure shows point estimates and 95-percent confidence intervals of plant-specific markdowns on age category indicators, controlling for indicators for plant size and industry, as well as state and year fixed effects. The omitted group is the smallest age category, less than three years, so coefficients reflect deviations relative to this baseline. Standard errors are clustered at the industry level.

Source: Authors' own calculations from ASM/CM data in 1976–2014.

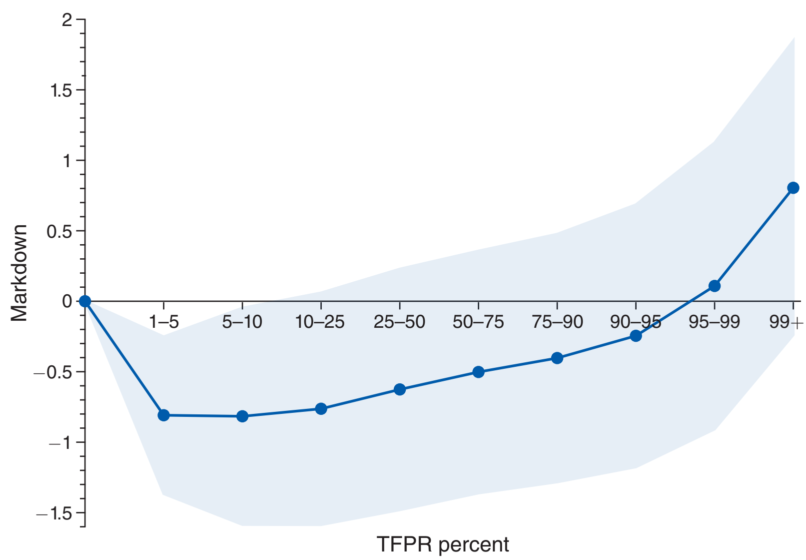

Figure 3. Markdown-TFPR Relationship

Notes: Nonparametric regressions of markdowns on productivity (employment-weighted). To avoid collinearity issues, we follow Haltiwanger, Jarmin, and Miranda (2013) and apply the normalization $\beta_1^{TFPR} = 0$ (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.

increasing in TFPR only after about the tenth percentile in the TFPR distribution. Though the variation in markdowns across the TFPR distribution is large, the coefficients on the productivity percentile categories are noisily estimated, and most estimates are not significantly different from zero at the 5 percent level.[29] Although there is suggestive evidence that the most productive establishments may have larger markdowns, because of the relative imprecision in these estimates we do not make strong inferences on the productivity-markdown relationship.

Scope and High-Tech Status.—Another feasible dimension of heterogeneity is the extent to which markdowns vary for plants belonging to firms with multiple establishments or with wide industrial or geographical scope. Such plants likely belong to firms with greater capitalization and internal networks for resource reallocation, potentially increasing the scope for markdowns (Giroud and Mueller 2019). We thus create binary variables that equal one when a plant is owned by a firm that has at least two active establishments ("multi-unit"), owned by a firm with establishments in two or more different six-digit NAICS industries ("multi-sector"), or owned by a firm with establishments in two or more counties ("multi-county").

Table 3 shows that plants owned by multi-unit firms charge markdowns more than 0.25 greater, on average, than stand-alone plants. We find quantitatively similar markdown premia for plants with greater industrial or geographical scope. These results continue to hold if we control for firm size.

We additionally investigate whether plants in the high-tech sector have higher markdowns.[30] 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 average markdowns for high-tech plants are (weakly) *lower* than for other plants, which is consistent with the results on more innovative firms in Van Reenen (1996).

Heterogeneous Labor.—Our baseline estimates allow for different types of labor across plants but implicitly assume that labor is homogeneous *within* plants. The ASM and CM break down the plant wage bill into components of production and nonproduction workers, allowing us to test for heterogeneity in markdowns across these two types of labor (treating them as separate inputs in the production function).[31] Table 4 shows estimated markdowns by industry separately for each labor type.

Allowing for labor heterogeneity does not greatly affect the pattern from our original estimates. Markdowns for nonproduction workers correspond closely to the baseline in Table 1, while those for production workers, while more variable, are

[29] The U-shape relationship between markdowns and TFPR does not appear to be driven by outliers in TFPR and/or entering and exiting plants. If we drop these observations from our sample, the U-shape flattens somewhat, but neither the qualitative pattern nor statistical significance changes.

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

[31] The US Census Bureau defines production workers as those "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.

Table 3—Firm's Characteristics and Markdowns

	Markdowns			
	Multi-unit	Multi-sector	Multi-county	High-tech
Premium	0.2514	0.2543	0.2558	−0.09054
	(0.04236)	(0.04173)	(0.04247)	(0.1081)
Observations (in millions)	1.393	1.393	1.393	1.393
R^2	0.2668	0.2696	0.2697	0.2511

Notes: See note to table 1 on markdown estimation. "Multi-sector" and "Multi-county" refers to a plant that is owned by a firm active in multiple six-digit NAICS industries or five-digit FIPS counties, respectively. A plant is considered "high-tech" based on its four-digit NAICS code and the categorization of Decker et al. (2016). Standard errors, in parentheses, are clustered at the industry level.

Source: Authors' calculations from ASM/CM data in 1976–2014.

somewhat higher, on average, than the baseline. However, there is little evidence of any systematic difference between the two groups that would suggest markdowns are driven by only one type of labor.[32]

That markdowns are similar for production and nonproduction labor may seem surprising to the extent that these groups are presumed synonyms for low-skill and high-skill, respectively, and that low-skill workers should have an easier time finding a comparable outside employment option. However, upon reflection, the pattern we find should not be surprising. First, production and nonproduction workers are *not* synonyms for low- and high-skill workers; rather, the former group, in addition to "fabricating, processing, [and] assembling," also includes highly skilled craftspersons, inspectors, and product developers.

Second, the summary results in Table 4 subsume spatial heterogeneity. The portability of a worker's skills across jobs depends not only on their type of skill but on the demand for that skill in the local labor market.[33] Thus, it is plausible that production workers are subject to lower markdowns in some locations (where the opportunities for alternative employment are plentiful) but not in others (where alternative employers are scarce), and this accords with their greater dispersion in markdowns. Third, it is not clear a priori whether production workers are more subject to labor market power than nonproduction workers. Outside employment options for both groups may be limited by noncompete employment contracts (Starr, Prescott, and Bishara 2021), which are quite prevalent in manufacturing (Colvin and Shierhold 2019). Indeed, using a structural nested logit model, Azar, Berry, and Marinescu (2019) find little difference in labor market power between higher- and lower-paying occupations.

[32] It is also reassuring that, even under the strict interpretation of Assumption VI and Proposition 1, we continue to find markdowns for production workers.

[33] Marinescu and Rathelot (2018) show that 81 percent of job seekers apply within their metropolitan area of residence, while Macaluso (2019) finds that earnings of laid-off workers recover faster if their last job used skills common to many jobs in the workers' metropolitan area.

TABLE 4—MARKDOWNS FOR PRODUCTION AND NONPRODUCTION WORKERS

	Panel A. Nonproduction		*Panel B. 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
Nonmetallic 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	1.530	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
Whole sample	1.682	1.488	1.963	1.527
Baseline	1.530		1.364	

Notes: See note to Table 1 on markdown estimation. The summary statistics under panels A and B reflect markdowns applied to nonproduction and production workers, respectively.

Source: Authors' calculations from ASM/CM data in 1976–2014.

III. Secular Trends in Aggregate Market Power

A. *Aggregation of Markdowns*

Thus far, we have focused on cross-sectional markdown dispersion, pooling across years, and have shown that (i) the average manufacturing plant operates in a monopsonistic environment and, (ii) plant-level markdowns vary substantially across and within industries but are positively associated with plant size. While an increase in labor market power is consistent with several observed secular trends in the US economy, there is still little direct time-series evidence for widening gaps between marginal revenue product of labor and wages (Syverson 2019). In this section, we investigate time trends in *aggregate* markdowns to gauge whether monopsony in US manufacturing has increased over time.

Although we have estimates for markdowns at the plant level, aggregation is not straightforward. Previous studies on markups have relied on weighted averages based on sales (De Loecker, Eeckhout, and Unger 2020) or employment (Rossi-Hansberg, Sarte, and Trachter 2020), but it is unclear in which context and for which questions it is appropriate to use these particular weights for markdown aggregation.[34] We

[34] Aggregation is more straightforward when one is willing to impose more structure. Berger, Herkenhoff, and Mongey (2022) show that in their model, a labor market counterpart to Atkeson and Burstein (2008), Herfindahl indices of payroll are sufficient statistics to calculate aggregate labor market power, but this need not hold more generally.

propose instead a flexible measure of aggregate markdowns that is (1) theoretically consistent with aggregate wedges, in the spirit of Edmond, Midrigan, and Xu (2021), and (2) accounts for the local nature of labor markets.

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.[35] Hsieh and Klenow (2009) and Itskhoki and Moll (2019) use similar approaches in defining aggregate productivity as a function of micro-level productivities. Importantly, we do not have to impose a specific structure for labor or output markets in order to achieve consistency with aggregate wedges. Consequently, our measure for the aggregate markdown is consistent with a variety of monopsony models.[36]

Second, several studies have shown that labor markets are "local" because workers find it costly to search for jobs in locations far from where they reside. For instance, Manning and Petrongolo (2017) estimate that the attractiveness of jobs decays sharply with distance, while Marinescu and Rathelot (2018) find that job seekers are 35 percent less likely to apply to a job 10 miles away from their zip code of residence. It is similarly costly to search in settings using different skills or performing different tasks (Kambourov and Manovskii 2009).

We thus characterize a *local* labor market as a sector-location pair, using three-digit NAICS codes and counties, resulting in a total of more than 20 distinct sectors (within manufacturing) and over 3,000 locations. In what follows, we denote sectors by j and locations by l.[37]

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.[38] We define the aggregate markdown $\mathcal{V}_{jlt}$ as the 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:

$$\frac{\theta^L_{jlt}}{\alpha^L_{jlt}} = \mathcal{M}_{jlt} \cdot \mathcal{V}_{jlt} \tag{10}$$

$$\frac{\theta^M_{jlt}}{\alpha^M_{jlt}} = \mathcal{M}_{jlt}, \tag{11}$$

where, with some abuse of notation, θ^L_{jlt} and α^L_{jlt} are, respectively in some market, the aggregate output elasticity of labor and the labor share. These objects are defined analogously for material inputs. We say that any measures for the aggregate markup $\mathcal{M}_{jlt}$ and markdown $\mathcal{V}_{jlt}$, that are based on micro-level markups and markdowns, are consistent with aggregate wedges whenever $\mathcal{M}_{jlt}$ and $\mathcal{V}_{jlt}$ satisfy equations (10) and (11). Then, we can show the following:

[35] 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); Galí, Gertler, and Lopez-Salido (2007); and Karabarbounis (2014). In particular, the aggregate wedge that defines the aggregate markdown in our setup is part of the gap between marginal products of labor and real wages in Karabarbounis (2014).

[36] We discuss these in online Appendix O.8.

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

[38] Edmond, Midrigan, and Xu (2021) adhere to a similar definition but instead assume that labor is fully flexible.

PROPOSITION 2: *Let Assumption I hold. Furthermore, let Assumptions II–VI hold for material inputs and Assumptions II and IV–VI hold for labor. If firm-level wage schedules are differentiable, then the* ***aggregate markdown*** *and* ***aggregate markup*** *for a local labor market (j,l) are consistent with aggregate wedges whenever they are equal to*

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

$$\mathcal{M}_{jlt} = \left(\sum_{i \in F_t(j,l)} s_{it} \cdot \frac{\theta_{it}^M}{\theta_{jlt}^M} \cdot \mu_{it}^{-1}\right)^{-1}, \tag{13}$$

where s_{it} are sales weights (i.e., $s_{it} = \frac{p_{it}y_{it}}{P_{jlt}Y_{jlt}}$) and $F_t(j,l)$ denotes the set of firms in labor market (j,l).

PROOF:

See Appendix B.

Whenever the market for material inputs is perfectly competitive, we can use an insight similar to the one used in Proposition 1. Recall that 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, as in equation (13), similar to Edmond, Midrigan, and Xu (2021).

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

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

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. If production technologies are not Cobb-Douglas, we need only apply correction terms that deal with heterogeneity in output elasticities—as can be seen from equations (12) and (13) in Proposition 2.

We then aggregate across labor markets through employment weights, defining the aggregate markdown as

$$\mathcal{V}_t = \sum_{j \in J} \sum_{l \in L} \omega_{jlt} \mathcal{V}_{jlt}, \tag{14}$$

where $\mathcal{V}_{jlt}$ is as in (12), and ω_{jlt} denotes the employment share of labor market (j,l). Following the literature on markups (e.g., De Loecker, Eeckhout, and Unger 2020)

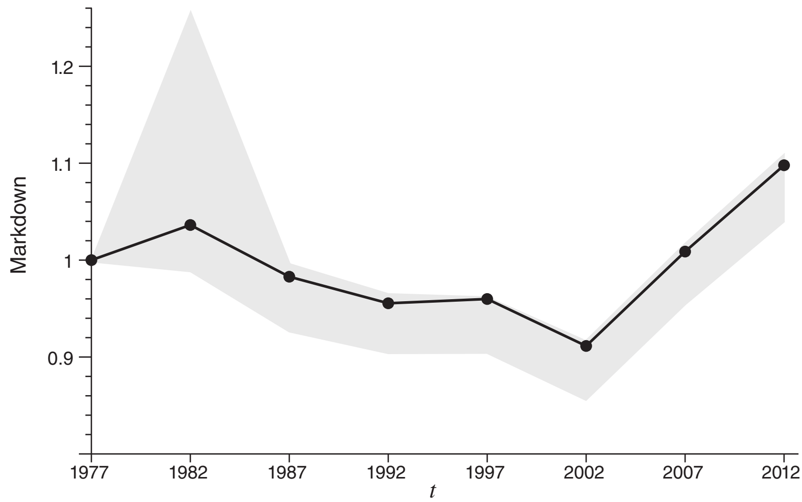

Figure 4. Time Evolution of the Aggregate Markdown, 1977–2012

Notes: Markdowns are constructed under the assumption of translog production and aggregated according to expressions (12) and (14). The aggregate markdown is normalized relative to its initial value in 1977. Standard errors are obtained through a block bootstrap procedure but are not symmetric because production function parameters enter firm-level markdowns in a highly nonlinear fashion, and firm-level markdowns also enter the aggregate markdown nonlinearly.

Source: Authors' own calculations from quinquennial CM data from 1977–2012.

and concentration (Autor et al. 2020; Rossi-Hansberg, Sarte, and Trachter 2020), we proceed by constructing markdowns at the firm, rather than plant, level using the CM.[39]

Figure 4 illustrates the resulting time trend of aggregate markdowns, $\mathcal{V}_t$. In contrast with previous trend estimates of markups (e.g., De Loecker, Eeckhout, and Unger 2020), the aggregate markdown $\mathcal{V}_t$ is not monotonic. Instead, $\mathcal{V}_t$ falls between the early 1980s and early 2000s, after which it begins to sharply increase. This pattern is inconsistent with 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. Yet, the stark increase in the aggregate markdown since this time is interesting, as others have noted acceleration in the decline in US business dynamism over the same horizon (see, e.g., Decker et al. 2016).[40]

[39] By construction, the aggregate markdown is an employment-weighted average of markdowns at the market level. The latter is constructed using equations (12) and (13). 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. This is sufficient for use in a repeated cross-section, as in our earlier analyses of the distribution of plant-level markdowns, but not for employment-weighted aggregation. In particular, the number of observations available to construct $\mathcal{V}_{jlt}$ and $\mathcal{M}_{jlt}$ might be rather small for some labor markets (j, l) and induce measurement error biases in these objects. Thus, we instead utilize the CM, which contains the universe of manufacturing plants but only at a quinquennial frequency.

[40] To understand which groups of firms determine movements in the aggregate markdown, we have applied a decomposition in the spirit of Foster, Haltiwanger, and Krizan (2001) to $\mathcal{V}_{jlt}$. This decomposition analyzes the role of changes within firms, across firms, and through firm entry and exit. As documented in online Appendix O.2, no single component drives the trend. In online Appendix O.2, we show that the trend in the aggregate markdown also

Contrasting the time series for the aggregate markdown in (14) 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 a labor market equivalent of the aggregate markup measure used in De Loecker, Eeckhout, and Unger (2020):

$$\begin{aligned} \mathcal{V}_t^{\mathrm{dLEU}} &= \sum_{p \in P_t} \omega_{pt} \nu_{pt} \\ &= \sum_{f \in F_t} \omega_{ft} \left[\sum_{p \in P_t(f)} s_{pft} \nu_{pt} \right] \\ &\equiv \sum_{f \in F_t} \omega_{ft} \nu_{ft}\,, \end{aligned} \tag{15}$$

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, $\mathcal{V}_t^{\mathrm{dLEU}}$ 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.

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 (2020). This approach still aggregates micro-level markdowns through employment weights, similarly to equation (15), 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 to

$$\mathcal{V}_t^{\mathrm{RHST}} = \sum_{j \in J} \sum_{l \in L} \omega_{jlt} M_{jlt} \quad \text{with} \quad M_{jlt} = \sum_{f \in F_t(j,l)} \omega_{fjlt} \nu_{fjlt}. \tag{16}$$

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 $\mathcal{V}_t^{\mathrm{dLEU}}$ follows our measure in a qualitative sense, $\mathcal{V}_t^{\mathrm{RHST}}$ monotonically decreases over the whole period.[41]

B. *Comparing Markdowns with Concentration Indices*

Several recent studies have used measures of concentration—either in output or input markets—as proxies for market power, both cross-sectionally and longitudinally. In this subsection, we discuss whether concentration is an accurate proxy

cannot be explained by changes in the composition of local labor markets or excluding health and pension benefits from the labor share.

[41] These differing trends can be rationalized by how markdowns are aggregated at the market level. Markdowns are aggregated linearly under the measure $\mathcal{V}_{jlt}^{\mathrm{RHST}}$, but $\mathcal{V}_{jlt}$ is constructed through the ratio of two harmonic weighted averages. Furthermore, each of these averages contain markups and reflect heterogeneity in output elasticities for labor and material inputs. Empirically, we have confirmed these latter factors explain more of the difference between $\mathcal{V}_t^{\mathrm{RHST}}$ and $\mathcal{V}_t$.

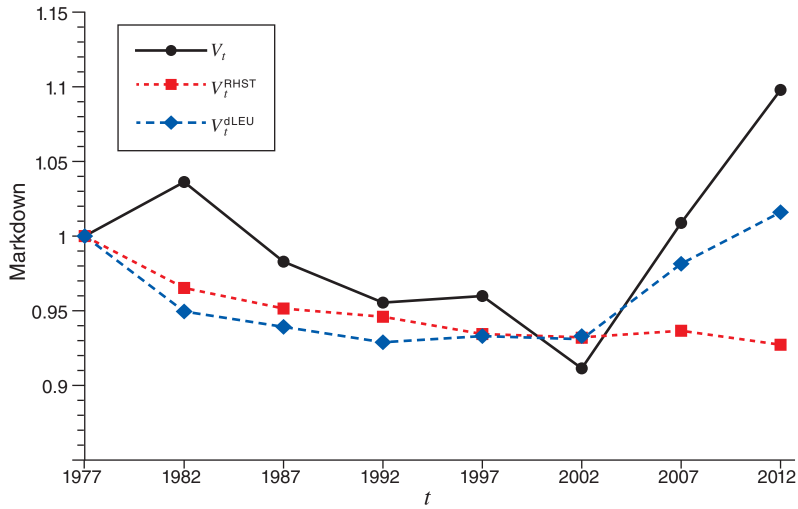

Figure 5. Trend Comparison of Different Aggregate Markdown Measures, 1977–2012

Notes: Markdowns are constructed under the assumption of translog production and aggregated according to expressions (14) for $\mathcal{V}_t$ (solid black), (15) for $\mathcal{V}_t^{\mathrm{dLEU}}$ (dashed blue), and (16) for $\mathcal{V}_t^{\mathrm{RHST}}$ (dashed red), respectively. All measures are normalized relative to their initial value in 1977.

Source: Authors' own calculations from quinquennial CM data from 1977–2012.

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 (Autor et al. 2020; Rossi-Hansberg, Sarte, and Trachter 2020) and has been increasingly popular in studies of labor markets, as well (Azar et al. 2020; Azar, Marinescu, and Steinbaum 2020; Benmelech, Bergman, and Kim 2020; Dodini et al. 2020; Rinz 2020). Yet, there is no a priori reason concentration and market power must be positively correlated. It may seem intuitive that large employers are able to exert more labor market power, but as Syverson (2019) points out for output markets, a negative correlation can arise naturally in the framework of Melitz and Ottaviano (2008) and has been empirically observed in several studies (Syverson 2004a; Syverson 2004b; Goldmanis et al. 2010). Despite these critiques, concentration indices have never been explicitly compared to direct, wedge-based measures of market power, at least not at a scale as wide as the whole manufacturing sector. This is precisely our aim 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:

$$\mathrm{HHI}_{mt} = \sum_{f \in F_t(m)} \left(\frac{x_{ft}}{X_{F(m)t}} \right)^2 \quad \text{s.t.} \quad X_{F(m)t} = \sum_{f' \in F_t(m)} x_{f't}, \tag{17}$$

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 focus on labor markets and thus set $m = (j, \ell)$ to remain consistent with our previous analyses. By construction, the

HHI ranges from $1/F_t(m)$ to 1. A value of 1 indicates maximum concentration—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.

There are two common approaches to combining market-level concentration measures into an aggregate measure. Under the first approach, HHIs are constructed at the industry level (so that a market m is a national industry) and then aggregated through employment or sales weights. Following Autor et al. (2020), we refer to these as measures of national concentration.

In contrast to this "national" approach, Rossi-Hansberg, Sarte, and Trachter (2020) have argued that market competition is sometimes better captured at the local level, which may especially be the case for labor. Therefore, markets are instead defined through sector-location cells. Formally,

$$\text{LOCAL}_t = \sum_{j \in J} \sum_{l \in L} \omega_{jlt} \text{HHI}_{jlt}. \tag{18}$$

Following our reasoning on the local nature of labor markets as in Section IIIA, we implement (18) with data on employment as our preferred measure underlying both x_{flt} and ω_{jlt}, where the latter are sector-location cell shares of total employment.[42]

Before we turn to comparisons of aggregates, we first correlate HHIs and our measure of markdowns across local labor markets (sector-location cells). We find that the cross-sectional correlation between $\mathcal{V}_{jlt}$ and HHI_{jlt} is essentially zero: across years, this correlation never exceeds 0.02, and it is sometimes negative.[43] Despite this weak cross-sectional correlation, Figure 6 demonstrates that time trends in *aggregate* local concentration (LOCAL_t) and markdowns ($\mathcal{V}_t$) are qualitatively similar. Nonetheless, while both series generally decline between the late 1970s and early 2000s, the subsequent rise in the aggregate markdown occurs both sooner and faster than the uptick in employment concentration.

These patterns suggest that—at least within manufacturing—cross-sectional and temporal variation in local employment concentration may not necessarily reflect variation in employer market power as measured by markdowns. While it is beyond the scope of this paper to thoroughly analyze these differences, we believe a promising area of future research is advancing theory on the general conditions under which measures of concentration serve as sufficient statistics for wedges between wages and marginal revenue products of labor.[44]

[42] Rossi-Hansberg, Sarte, and Trachter (2020) focus on product markets and apply the analogue of equation (18) to sales for x_{jlt} but employment for ω_{jlt}. Rinz (2020), like us, focuses on labor markets and uses employment for both x_{jlt} and ω_{jlt}, but uses the Longitudinal Business Database to cover all establishments, not just those in manufacturing. While our results here consider (stock) employment concentration, we have also constructed concentration measures based on vacancies (as in Azar et al. 2020), job creation flows, and payroll, all of which produce qualitatively similar patterns. Results for vacancies, based on data from Burning Glass Technologies (BGT), can be found in online Appendix O.9. Remaining results are available upon request.

[43] We provide full details in online Appendix O.2.

[44] The weak empirical relationship we document may stem, in part, from different definitions of the relevant labor market. Azar, Marinescu, and Steinbaum (2019), for example, find a negative correlation between job application elasticity and HHI when markets are defined by occupation-commuting zone pairs. Data limitations unfortunately prevent us from investigating the relative roles of occupation versus industry, or sectoral composition, in explaining these differences.

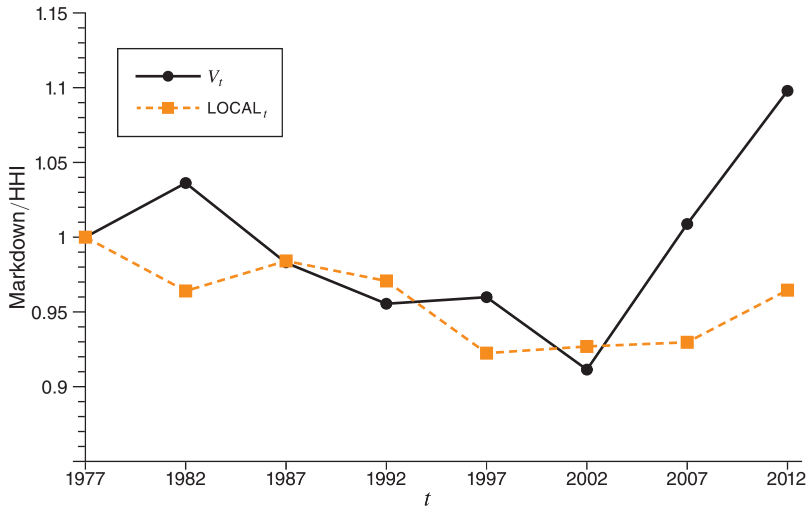

Figure 6. Aggregate Markdown and Local Concentration, 1977–2012

Notes: The solid black line shows the time series for the aggregate markdown as in (14), and the dashed orange line shows the time series of local employment concentration as in (18). Both are normalized to their initial respective values in 1977.

Source: Authors' own calculations from quinquennial CM data from 1977–2012.

IV. Robustness

In this section, we discuss the robustness of our results. We present several exercises that address concerns related to the validity of our markdown formula in Proposition 1. We also discuss some unresolved econometric issues of the proxy variable methodology.

A. *Choice of Flexible Input*

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 is the requirement for at least one flexible input. Pinpointing such an input is difficult in most publicly available datasets, as most inputs are not observed separately but rather aggregated into broad groups following accounting standards.[45] Although there is still some disagreement 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 (2020), using transaction-level

[45] Recent studies estimating markups typically rely on the Compustat database, in which variable inputs are often identified with "cost of goods sold" (COGS)—which commingles material inputs and variable and fixed labor—or "selling, general, and administrative expenses" (SGA). Because our data allow us to observe *separately* expenditures on capital, labor, material, and energy, we circumvent having to make this choice. Regrettably, neither data source—or any other with similar coverage, to our knowledge—further allows for observation of input quality or other sources of heterogeneity.

data from French manufacturers, finds evidence of market power in imported intermediate inputs under the identifying assumption that domestically sourced intermediate inputs are 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 (3) 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, ν_M implicitly exceeds unity and our estimates for labor markdowns would be biased toward zero, underestimating the extent of labor monopsony in US manufacturing.

A plausible alternative for the flexible input is energy, as advocated by Kim (2017). He maintains that monopsony power through buyer-supplier networks may affect 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 the energy input market is perfectly competitive. They find that plant-level differences within manufacturing industries in energy purchases account for a substantial fraction—at least one-third—of overall price dispersion. Furthermore, they document sizable price gaps between larger and smaller purchases, even when controlling for plant location and/or electric utility provider fixed effects. This seems to contradict the "no monopsony" condition for energy and cautions against its use as a flexible input. Moreover, material inputs have another attractive property in that they represent a much larger share of manufacturing revenues than does energy. Because our measure of markdowns requires division by the flexible input, measurement error is of lesser concern for material inputs compared to energy.[46] We view these factors as compelling evidence in favor of materials as the flexible input.

B. *Point Identification*

Gandhi, Navarro, and Rivers (2020) show that the standard assumptions of the proxy variable method (as we describe in Subsection IB) are insufficient to point-identify production function parameters, and that additional sources of variation in the demand for flexible inputs are required. In turn, Flynn, Gandhi, and Traina (2019) have shown that point identification can be restored if the returns to scale of the production function is known. They suggest that constant returns to scale is a useful benchmark that performs well in their Monte Carlo simulations. When we impose this constant returns to scale assumption in our context, we reassuringly find our markdown estimates change relatively little (column "CRS" of Table 5).[47] This corroborates the notion that our strategy yields reliable estimates of monopsony power in US manufacturing.

[46] We provide evidence of both factors in online Appendix O.1. If we calculate markdowns using energy as the flexible input, we find higher markups and lower markdowns, suggesting labor markdowns are low *relative* to energy markdowns. Additionally, labor markdowns calculated with energy as the flexible input have volatility nearly an order of magnitude greater than our baseline estimates, suggesting division bias.

[47] Additional details of this exercise are in online Appendix O.3.

Table 5—Robustness of Markdown Estimates

Industry group	Baseline (1)	CRS (2)	Biennial (3)	Benefits (4)
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
Nonmetallic 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

Notes: 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 column 2, we display estimates under the additional assumption of constant returns to scale to address identification concerns. Results from estimating markdowns using biennial data to capture nonconvex adjustment costs are displayed under column 3. Results from including benefits in the measure of labor compensation (available only from 2002 forward) are displayed under column 4.

Source: Authors' calculations from ASM/CM data in 1976–2014.

C. *Labor Adjustment Costs*

In our baseline specification, we assumed there are no labor adjustment costs (Assumptions II and IV). 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 expressions 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 trivially affect our baseline estimates, we proceed in two steps. First, we show that, when labor is subject to convex adjustment costs, the wedge between the marginal revenue product of labor and wages reflects both monopsony power and adjustment costs. In particular, we show that $\frac{R'(\ell^*)}{w(\ell^*)} = \left(\varepsilon_S^{-1} + 1\right) + \mathcal{A}$, 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 (Hall 2004; Cooper, Haltiwanger, and Willis 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 correction adjusts average markdowns by approximately 0.03, quite small relative to the baseline average markdown of 1.53.[48]

[48] We provide derivations of the correction term and a detailed illustration of this exercise in online Appendix O.4.

We also consider the possibility of fixed or otherwise *non*convex adjustment costs by reestimating our markdowns on a biennial basis. Conceptually, nonconvex adjustment costs that may affect our estimates at an annual frequency are less likely to do so at a 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 5 under the column "Biennial," are again similar to baseline.

D. *Benefits*

Our baseline measure of labor costs is based on "wages" and covers a broad range of compensation, including base salaries and wages; bonuses; incentive, overtime, and shift differential pay; and stock grants and options. However, it does not include employer-provided 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 are correlated with the components of markdowns. We thus reestimate micro-level markdowns including benefits in our compensation measure. This more inclusive measure of labor compensation is available only from 2002 onward, which results in a smaller sample than our baseline estimates.[49] The results, displayed in the last column of Table 5, show markdown estimates that are slightly lower than baseline, indicating that the inclusion of benefits may be a nontrivial part of the wedge between observed wages and marginal revenue product of labor. Nonetheless, these estimated markdowns are still above unity for each industry, demonstrating that the presence of monopsony is robust to broader measures of compensation.

E. *Revenue versus Quantity Elasticities*

In a recent paper, Bond et al. (2021) argue that the production approach and its implementation with proxy variable estimators cannot generate unbiased estimators of market power. Their critique centers around three issues that we discuss here. The first is that, in most firm-level datasets, the quantity of physical output is not available and instead has to be proxied by deflating revenues. As noted by Klette and Griliches (1996), this can lead to downward-biased estimates for markups. Bond et al. (2021) further argue that markups under the production approach will mechanically equal unity whenever (deflated) revenues are used to proxy for physical output.

We argue that this is indeed problematic when markups need to be identified in isolation. However, this critique does not apply to markdowns. The key insight is that markdowns are estimated through a *ratio* of elasticities. Revenue elasticities are not equal to output elasticities; however, the component (or "bias") that separates them is identical across inputs and multiplicative. As a result, the bias documented by Bond et al. (2021) cancels out via our construction of markdowns: the ratio of revenue

[49] Online Appendix O.5 provides full details on the components of labor compensation in the ASM/CM data.

elasticities for two inputs is equal to the ratio of output elasticities for these two same inputs. We formalize this intuition in Proposition 4 of online Appendix O.6.

F. *Inputs for Nonproduction Purposes*

Bond et al. (2021) also argue that the production approach can lead to biased estimates of markups and markdowns when the econometrician cannot separate inputs used for purposes other than production that could still affect the quantity of output. For example, inputs could also be used to shift demand (e.g., marketing/advertising). To ensure that our estimates are not subject to this criticism, we need to show that material inputs and labor are used primarily for production purposes.

We argue that it is unlikely *prima facie* that material inputs are used to influence demand. Note that these inputs consist of raw materials, parts, containers, and supplies. Given this categorization, it is safe to assume that material inputs are used solely for production purposes.

However, it is less obvious that no labor inputs are used for shifting demand. We perform two robustness exercises to address this possibility. First, as noted in Section IIB, our data allow us to separate labor into production and nonproduction workers, and when we estimate markdowns for these types of labor separately, we find that monopsony forces are still significant among production workers, specifically. Second, and more generally, our focus on manufacturing *plants* should render our analysis more robust to this specific critique, since we can plausibly assume that the great majority of a manufacturer's plant-level work force is indeed employed for production. In fact, we explicitly derive a markdown counterpart for the bias characterized in Bond et al. (2021) and show that, even if we do not separate production from nonproduction labor, the components inducing bias are likely to be small for manufacturers.[50]

G. *Scalar Unobservable Assumption*

The last critique in Bond et al. (2021) relates to the scalar unobservable assumption (our Assumption 4). They show that, in the presence of market power, this assumption cannot be satisfied since the econometrician is also required to observe a plant's marginal cost of production. Consequently, they suggest using production function estimators that do *not* rely on the scalar unobservable assumption, such as dynamic panel instrumental variable methods (Blundell and Bond 2000). To evaluate this claim, we use Monte Carlo methods to compare the performance of several production function estimators. In particular, we adopt data-generating processes from Ackerberg, Caves, and Frazer (2015) that are inconsistent with the econometric assumptions of the family of proxy variable estimators. Nevertheless, as we show in online Appendix O.6, our preferred translog estimator outperforms several estimators that do not rely on the scalar unobservable assumption, including those from Blundell and Bond (2000) and Hu, Huang, and Sasaki (2020). Hence, even though the scalar unobservable assumption is violated, we do not believe it causes significant problems in practice.

[50] The details for this exercise are found in online Appendix O.6.

V. Conclusion

This paper provides a characterization of employer market power in the US 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 US 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. For our whole sample, the interquartile range of markdowns is 0.618, but most of this variation is observed within detailed industries, with an average within-industry interquartile range of 0.616. Furthermore, we find that size—whether measured as the relative share of employment in a local labor market or as geographical and sectoral scope—is associated with greater markdowns. On the other hand, we find less correlation with a revenue-based measure of productivity or an indicator for being in a high-tech industry.

We also investigate long-term trends in employer market power, via a novel measure of aggregate markdowns that is consistent with aggregate wedges, accounts for local labor markets, and uses sales-weighted harmonic averages to adjust for production heterogeneity across firms. We find aggregate markdowns decreased between the late 1970s and early 2000s but increased sharply afterward. This nonmonotonic pattern is inconsistent with the view that the decline in the US labor share (or wage stagnation) was 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 that the variation underlying local employment concentration does not necessarily reflect the variation underlying employer market power as measured by markdowns.[51]

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, while we provide qualitative evidence of a negative correlation between the industry's rate of unionization and markdowns, we do not yet know whether the cross-industry variation in markdowns can be further rationalized by the prevalence of noncompete agreements or labor regulations (e.g., right-to-work laws). 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

[51] Recent papers have documented that large increases in HHI driven by mergers lead to decreased wages (Arnold 2020; Prager and Schmitt 2021). It is unclear, though, whether this relationship holds throughout the HHI distribution, or whether the reduction in wages stems from labor market frictions other than the wedge between wages and marginal product.

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. Last, our econometric methodology does not explicitly allow for factor-biased technological change. While there are estimation methods that do account for labor-augmenting technological change, they do not allow for a generalized production function (Doraszelski and Jaumandreu 2018; Raval 2020) and/or labor market power (Demirer 2020). We leave investigation of these themes for future research.

Appendix A. Proof of Proposition 1

We start with the cost minimization problem of a firm. In general, we have

$$\text{(A1)} \qquad \min_{\mathbf{X}_{it}\in\mathbb{R}_+^K} \sum_{k=1}^{K} V_{it}^k\left(X_{it}^k\right) X_{it}^k + \Phi_t^k\left(X_{it}^k, X_{it-1}^k\right) \quad \text{s.t.} \quad F\left(\mathbf{X}_{it};\omega_{it}\right) \geq Q_{it},$$

where $\mathbf{X}_{it} = \left(X_{it}^1, \ldots, X_{it}^K\right)'$ is the firm's vector of $K > 1$ production inputs with prices $\left\{V_{it}^k\right\}_{k=1}^K$. Furthermore, ω_{it} denotes a firm i's productivity level at time t whereas a firm's production technology is denoted by $F\left(\mathbf{X}_{it};\omega_{it}\right)$. Adjustment costs for some input k are captured by the function $\Phi_t^k\left(\cdot,\cdot\right)$.

To derive markdowns, we start with the insight by Hall (1986) 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). Let Assumption I hold. Furthermore, let Assumptions II–VI hold for some input k'. Then, a firm i's markup satisfies

$$\text{(A2)} \qquad \mu_{it} = \frac{\partial F\left(\mathbf{X}_{it};\omega_{it}\right)}{\partial X_{it}^{k'}} \frac{X_{it}^{k'}}{Q_{it}} \cdot \left(\frac{V_{it}^{k'} X_{it}^{k'}}{P_{it} Q_{it}}\right)^{-1}$$

$$\equiv \frac{\theta_{it}^{k'}}{\alpha_{it}^{k'}}.$$

PROOF:

Under the stated assumptions, the first order condition for any flexible input k', associated with cost minimization problem (A1), satisfies

$$V_{it}^{k'} = \lambda_{it} \frac{\partial F\left(\mathbf{X}_{it};\omega_{it}\right)}{\partial X_{it}^{k'}},$$

where λ_{it} is the Lagrangian multiplier associated with the cost minimization problem in (A1). 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

$$\frac{V_{it}^{k'} X_{it}^{k'}}{P_{it} Q_{it}} = \frac{\lambda_{it}}{P_{it}} \frac{\partial F\left(\mathbf{X}_{it};\omega_{it}\right)}{\partial X_{it}^{k'}} \frac{X_{it}^{k'}}{Q_{it}},$$

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:

$$\mu_{it} = \frac{\theta_{it}^{k'}}{\alpha_{it}^{k'}}, \tag{A3}$$

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' that satisfies Assumptions II–VI is sufficient to establish this result. ∎

Given this lemma, we can prove the main result of Proposition 1. Without loss of generality, consider the following *conditional* cost minimization problem:

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

where $\mathbf{X}^*_{-\ell,it}$ denotes the vector of optimized inputs with the exception of labor ℓ_{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

$$\begin{aligned} \left[\frac{w'_{it}(\ell_{it})\,\ell_{it}}{w_{it}(\ell_{it})} + 1\right] &\equiv \varepsilon_S^{-1}(\ell_{it}) + 1 \\ &= \frac{\lambda_{it}}{P_{it}} \cdot \frac{\partial F(\ell_{it}, \mathbf{X}^*_{-\ell,it};\omega_{it})}{\partial \ell_{it}} \frac{\ell_{it}}{Q_{it}} \cdot \frac{P_{it} Q_{it}}{w_{it}(\ell_{it})\,\ell_{it}} \\ &\equiv \mu_{it}^{-1} \cdot \frac{\theta_{it}^{\ell}}{\alpha_{it}^{\ell}}. \end{aligned} \tag{A4}$$

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

$$\frac{\theta_{it}^{\ell}}{\alpha_{it}^{\ell}} = \nu_{it} \cdot \mu_{it}. \tag{A5}$$

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

$$\nu_{it} = \frac{\theta_{it}^{\ell}}{\alpha_{it}^{\ell}} \cdot \left(\frac{\theta_{it}^{k'}}{\alpha_{it}^{k'}}\right)^{-1}. \tag{A6}$$

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_{it}^{ℓ} and α_{it}^{M} 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.

Appendix B. Proof of Proposition 2

Whenever Assumptions I–VI are satisfied and Assumptions II–VI apply specifically to material inputs, then we showed in Lemma 1 that markups can be characterized as

$$\mu_{it} = \frac{\theta_{it}^M}{\alpha_{it}^M} = \theta_{it}^M \cdot \frac{P_{it} Q_{it}}{P_t^M M_{it}} \tag{B1}$$

Similar to Edmond, Midrigan, and Xu (2021), 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 (B1) also holds in the aggregate, i.e., we have

$$\mathcal{M}_t \equiv \theta_t^M \cdot \frac{P_t Y_t}{P_t^M M_t} \tag{B2}$$

where we dropped the indices for local markets (j, l) for simplicity. Substituting out the price for material inputs P_t^M from (B2) into (B1), we obtain

$$\mu_{it} = \frac{\theta_{it}^M}{\theta_t^M} \cdot \frac{P_{it} Q_{it}}{P_t Y_t} \cdot \frac{M_t}{M_{it}} \cdot \mathcal{M}_t.$$

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

$$\mathcal{M}_t = \left(\sum_{i \in F_t} \frac{\theta_{it}^M}{\theta_t^M} \cdot s_{it} \cdot \mu_{it}^{-1} \right)^{-1}, \tag{B3}$$

where $s_{it} \equiv \frac{P_{it} Q_{it}}{P_t Y_t}$ 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* $\mathcal{V}_t$. Whenever Assumptions II and IV–VI hold 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 1. Therefore, we have

$$\nu_{it} \mu_{it} = \theta_{it}^L \cdot \frac{P_{it} Q_{it}}{w_{it} \ell_{it}}. \tag{B4}$$

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

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

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

$$\mathcal{V}_t \cdot \mathcal{M}_t = \theta_t^L \cdot \frac{P_t Y_t}{w_t L_t}. \tag{B5}$$

Using our previous results, we then get

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

Apply expression (B3) for the aggregate markup and we obtain an expression for the aggregate markdown,

$$\mathcal{V}_t = \frac{\left(\sum_{i \in F_t} \frac{\theta_{it}^L}{\theta_t^L} \cdot s_{it} \cdot \left(\mu_{it} \nu_{it}\right)^{-1}\right)^{-1}}{\left(\sum_{i \in F_t} \frac{\theta_{it}^M}{\theta_t^M} \cdot s_{it} \cdot \mu_{it}^{-1}\right)^{-1}}. \tag{B6}$$

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

$$\mathcal{V}_t = \frac{\left(\sum_{i \in F_t} s_{it} \cdot \left(\mu_{it} \nu_{it}\right)^{-1}\right)^{-1}}{\left(\sum_{i \in F_t} s_{it} \cdot \mu_{it}^{-1}\right)^{-1}}, \tag{B7}$$

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

REFERENCES

Abowd, John M., Francis Kramarz, and David N. Margolis. 1999. "High Wage Workers and High Wage Firms." *Econometrica* 67 (2): 251–333

Ackerberg, Daniel A. 2020. "Timing Assumptions and Efficiency: Empirical Evidence in a Production Function Context." Unpublished.

Ackerberg, Daniel, Kevin Caves, and Garth Frazer. 2015. "Identification Properties of Recent Production Function Estimators." *Econometrica* 83 (6): 2411–51.

Arnold, David. 2020. "Mergers and Acquisitions, Local Labor Market Concentration, and Worker Outcomes." Unpublished.

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

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

Autor, David, David Dorn, Lawrence Katz, Christina Patterson, and John Van Reenen. 2020. "The Fall of the Labor Share and the Rise of Superstar Firms." *Quarterly Journal of Economics* 135 (2): 645–709.

Azar, José, Steven Berry, and Ioana Marinescu. 2019. "Estimating Labor Market Power." Unpublished.

Azar, José, Ioana Marinescu, and Marshall Steinbaum. 2019. "Measuring Labor Market Power Two Ways." *AEA Papers and Proceedings* 109: 317–21.

Azar, José, Ioana Marinescu, and Marshall I. Steinbaum. 2020. "Labor Market Concentration." *Journal of Human Resources* 57: S167–99.

Azar, José, Ioana Marinescu, Marshall Steinbaum, and Bledi Taska. 2020. "Concentration in US Labor Markets: Evidence from Online Vacancy Data." *Labour Economics* 66: 1018–86.
Baqaee, David, and Emmanuel Farhi. 2020. "Productivity and Misallocation in General Equilibrium." *Quarterly Journal of Economics* 135 (1): 105–63.
Basu, Susanto. 1995. "Intermediate Goods and Business Cycles: Implications for Productivity and Welfare." *American Economic Review* 85 (3): 512–31.
Benmelech, Efraim, Nittai K. Bergman, and Hyunseob Kim. 2020. "Strong Employers and Weak Employees: How Does Employer Concentration Affect Wages?" *Journal of Human Resources* 57: S200–50.
Berger, David, Kyle Herkenhoff, and Simon Mongey. 2022. "Labor Market Power." *American Economic Review* 112 (4): 1147–93.
Bhaskar, V., and Ted To. 1999. "Minimum Wages for Ronald McDonald Monopsonies: A Theory of Monopsonistic Competition." *Economic Journal* 109 (455): 190–203.
Blundell, Richard, and Stephen Bond. 2000. "GMM Estimation with Persistent Panel Data: An Application to Production Functions." *Econometric Reviews* 19 (3): 321–40.
Bond, Steve, Arshia Hashemi, Greg Kaplan, and Piotr Zoch. 2021. "Some Unpleasant Markup Arithmetic: Production Function Elasticities and Their Estimation from Production Data." *Journal of Monetary Economics* 121: 1–14.
Bontemps, Christian, Jean-Marc Robin, and Gerard J. Van den Berg. 1999. "An Empirical Equilibrium Job Search Model with Search on the Job and Heterogeneous Workers and Firms." *International Economic Review* 40 (4): 1039–74.
Brooks, Wyatt J., Joseph P. Kaboski, Illenin O. Kondo, Yao Amber Li, and Wei Qian. 2021a. "Infrastructure Investment and Labor Monopsony Power." *IMF Economic Review* 69 (3): 470–504.
Brooks, Wyatt J., Joseph P. Kaboski, Yao Amber Li, and Wei Qian. 2021b. "Exploitation of Labor? Classical Monopsony Power and Labor's Share." *Journal of Development Economics* 150: 102627.
Burdett, Kenneth, and Dale T. Mortensen. 1998. "Wage Differentials, Employer Size, and Unemployment." *International Economic Review* 39 (2): 257–73.
Card, David, Ana Rute Cardoso, Joerg Heining, and Patrick Kline. 2018. "Firms and Labor Market Inequality: Evidence and Some Theory." *Journal of Labor Economics* 36 (1): 13–70.
Card, David, Francesco Devicienti, and Agata Maida. 2014. "Rent-Sharing, Holdup, and Wages: Evidence from Matched Panel Data." *Review of Economic Studies* 81 (1): 84–111.
Chan, Mons, Kory Kroft, and Ismael Mourifie. 2019. "An Empirical Framework for Matching with Imperfect Competition." Unpublished.
Christofides, Louis N., and Andrew J. Oswald. 1992. "Real Wage Determination and Rent-Sharing in Collective Bargaining Agreements." *Quarterly Journal of Economics* 107 (3): 985–1002.
Cole, Harold, and Lee E. Ohanian. 2002. "The U.S. and U.K. Great Depressions Through the Lens of Neoclassical Growth Theory." *American Economic Review* 92 (2): 28–32.
Colvin, Alexander J.S., and Heidi Shierhold. 2019. *Noncompete Agreements*. Washington, DC: Economic Policy Institute.
Cooper, Russell, John Haltiwanger, and Jonathan L. Willis. 2007. "Search Frictions: Matching Aggregate and Establishment Observations." *Journal of Monetary Economics* 54 (S1): 56–78.
Davis, Steven J., Cheryl Grim, John Haltiwanger, and Mary Streitwieser. 2013. "Electricity Unit Value Prices and Purchase Quantities: U.S. Manufacturing Plants, 1963-2000." *Review of Economics and Statistics* 95 (4): 1150–65.
Decker, Ryan A., John Haltiwanger, Ron S. Jarmin, and Javier Miranda. 2016. "Declining Business Dynamism: What We Know and the Way Forward." *American Economic Review* 106 (5): 203–07.
De Loecker, Jan. 2011. "Recovering Markups from Production Data." *International Journal of Industrial Organization* 29 (3): 350–55.
De Loecker, Jan, Jan Eeckhout, and Gabriel Unger. 2020. "The Rise of Market Power and the Macroeconomic Implications." *Quarterly Journal of Economics* 135 (2): 561–644.
De Loecker, Jan, and Frederic Warzynski. 2012. "Markups and Firm-Level Export Status." *American Economic Review* 102 (6): 2437–71.
Demirer, Mert. 2020. "Production Function Estimation with Factor-Augmenting Technology: An Application to Markups." Unpublished.
Diamond, Peter A. 1982. "Wage Determination and Efficiency in Search Equilibrium." *Review of Economic Studies* 49 (2): 217–27.
Dobbelaere, Sabien, and Jacques Mairesse. 2013. "Panel Data Estimates of the Production Function and Product and Labor Market Imperfections." *Journal of Applied Econometrics* 28 (1): 1–46.
Dodini, Samuel, Michael Lovenheim, Kjell Salvanes, and Alexander Willén. 2020. "Monopsony, Skills, and Labor Market Concentration." Unpublished.

Doraszelski, Ulrich, and Jordi Jaumandreu. 2018. "Measuring the Bias of Technological Change." *Journal of Political Economy* 126 (3): 1027–84.
Edmond, Chris, Virgiliu Midrigan, and Daniel Yi Xu. 2021. "How Costly Are Markups?" Unpublished.
Eggertsson, Gauti, Jacob A. Robbins, and Ella Getz Wold. 2021. "Kaldor and Piketty's Facts: The Rise of Monopoly Power in the United States." *Journal of Monetary Economics* 124: S19–S38.
Elsby, Michael, Bart Hobijn, and Aysegul Şahin. 2013. "The Decline of the U.S. Labor Share." *Brookings Papers on Economic Activity* 43 (2): 1–52.
Flynn, Zach, Amit Gandhi, and James Traina. 2019. "Measuring Markups with Production Data." Unpublished.
Foster, Lucia, John Haltiwanger, and C.J. Krizan. 2001. "Aggregate Productivity Growth: Lessons from Microeconomic Evidence." In *New Developments in Productivity Analysis*, edited by Charles R. Hulten, Edwin R. Dean, and Michael J. Harper, 303–72. Chicago: University of Chicago Press.
Foster, Lucia, John Haltiwanger, and Chad Syverson. 2008. "Reallocation, Firm Turnover, and Efficiency: Selection on Productivity or Profitability?" *American Economic Review* 98 (1): 394–425.
Foster, Lucia, John Haltiwanger, and Chad Syverson. 2016. "The Slow Growth of New Plants: Learning about Demand?" *Economica* 83 (329): 91–129.
Gali, Jordi, Mark Gertler, and David J. Lopez-Salido. 2007. "Markups, Gaps, and the Welfare Costs of Business Fluctuations." *Review of Economics and Statistics* 89 (1): 44–59.
Gandhi, Amit, Salvador Navarro, and David A. Rivers. 2020. "On the Identification of Gross Output Production Functions." *Journal of Political Economy* 128 (8): 2973–3016.
Giroud, Xavier, and Holger M. Mueller. 2019. "Firms' Internal Networks and Local Economic Shocks." *American Economic Review* 109 (10): 3617–49.
Goldmanis, Maris, Ali Hortacsu, Chad Syverson, and Onsel Emre. 2010. "E-commerce and the Market Structure of Retail Industries." *Economic Journal* 120 (545): 651–82.
Griliches, Zvi, and Jacques Mairesse. 1998. "Production Functions: The Search for Identification." In *Econometrics and Economic Theory in the Twentieth Century: The Ragnar Frisch Centennial Symposium*, edited by S. Strom, 169–203. Cambridge, MA: Cambridge University Press.
Hall, Robert E. 1986. "Market Structure and Macroeconomic Fluctuations." *Brookings Papers on Economic Activity* 16 (2): 285–322.
Hall, Robert E. 1988. "The Relation between Price and Marginal Cost in U.S. Industry." *Journal of Political Economy* 96 (5): 921–47.
Hall, Robert E. 2004. "Measuring Factor Adjustment Costs." *Quarterly Journal of Economics* 119 (3): 899–927.
Haltiwanger, John, Ron S. Jarmin, and Javier Miranda. 2013. "Who Creates Jobs? Small versus Large versus Young." *Review of Economics and Statistics* 95 (2): 347–61.
Hsieh, Chang-Tai, and Peter J. Klenow. 2009. "Misallocation and Manufacturing TFP in China and India." *Quarterly Journal of Economics* 124 (4): 1403–48.
Hu, Yingyao, Guofang Huang, and Yuya Sasaki. 2020. "Estimating Production Functions with Robustness Against Errors in the Proxy Variables." *Journal of Econometrics* 215 (2): 375–98.
Itskhoki, Oleg, and Benjamin Moll. 2019. "Optimal Development Policies with Financial Frictions." *Econometrica* 87 (1): 139–73.
Jarosch, Gregor, Jan Sebastian Nimczik, and Issac Sorkin. 2021. "Granular Search, Market Structure, and Wages." Unpublished.
Kambourov, Gueorgui, and Iourii Manovskii. 2009. "Occupational Specificity of Human Capital." *International Economic Review* 50 (1): 63–115.
Karabarbounis, Loukas. 2014. "The Labor Wedge: MRS vs. MPN." *Review of Economic Dynamics* 17 (2): 206–23.
Karabarbounis, Loukas, and Brent Neiman. 2013. "The Global Decline of the Labor Share." *Quarterly Journal of Economics* 129 (1): 61–103.
Kehrig, Matthias. 2015. "The Cyclical Nature of the Productivity Distribution." Unpublished.
Kim, Ryan. 2017. "Price-Cost Markup Cyclicality: New Evidence and Implications." Unpublished.
Klette, Tor Jakob, and Zvi Griliches. 1996. "The Inconsistency of Common Scale Estimators When Output Prices are Unobserved and Endogenous." *Journal of Applied Econometrics* 11 (4): 343–61.
Lamadon, Thibaut, Magne Mogstad, and Bradley Setzler. 2022. "Imperfect Competition, Compensating Differentials and Rent Sharing in the U.S. Labor Market." *American Economic Review* 112 (1): 169–212.
Levinsohn, James A., and Amil Petrin. 2003. "Estimating Production Functions Using Inputs to Control for Unobservables." *Review of Economic Studies* 70 (2): 317–40.
Lipsius, Ben. 2018. "Labor Market Concentration Does Not Explain the Falling Labor Share." Unpublished.

Macaluso, Claudia. 2019. "Skill Remoteness and Post-Layoff Labor Market Outcomes." Unpublished.
Manning, Alan. 2003. *Monopsony in Motion.* Princeton, NJ: Princeton University Press.
Manning, Alan. 2011. "Imperfect Competition in the Labor Market." In *Handbook of Labor Economics*, Vol. 4, edited by O. Ashenfelter and David Card, 973–1041. North-Holland: Elsevier.
Manning, Alan, and Barbara Petrongolo. 2017. "How Local Are Labor Markets? Evidence from a Spatial Job Search Model." *American Economic Review* 107 (10): 2877–2907.
Marinescu, Ioana, and Roland Rathelot. 2018. "Mismatch Unemployment and the Geography of Job Search." *American Economic Journal: Macroeconomics* 10 (3): 42–70.
Melitz, Marc J., and Giancarlo I.P. Ottaviano. 2008. "Market Size, Trade, and Productivity." *Review of Economic Studies* 75 (1): 295–316.
Morlacco, Monica. 2020. "Market Power in Input Markets: Theory and Evidence from French Manufacturing." Unpublished.
Mortensen, Dale. 2003. "How Monopsonistic is the (Danish) Labor Market?" In *Knowledge, Information and Expectations in Modern Macroeconomics*, edited by Philippe Aghion, Roman Frydman, Joseph Stiglitz and Michael Woodford, 285–308. Princeton, NJ: Princeton University Press.
Mortensen, Dale T., and Christopher A. Pissarides. 1994. "Job Creation and Job Destruction in the Theory of Unemployment." *Review of Economic Studies* 61 (3): 397–415.
Olley, Steven, and Ariel Pakes. 1996. "The Dynamics of Productivity in the Telecommunications Industry." *Econometrica* 64 (6): 1263–97.
Pesaran, M. Hashem, and Ron Smith. 1995. "Estimating Long-Run Relationships from Dynamic Heterogeneous Panels." *Journal of Econometrics* 68 (1): 79–113.
Posner, Eric A., Geyl Weyl, and Suresh Naidu. 2018. "Antitrust Remedies for Labor Market Power." *Harvard Law Review* 132 (2): 536–601.
Prager, Elena, and Matt Schmitt. 2021. "Employer Consolidation and Wages: Evidence from Hospitals." *American Economic Review* 111 (2): 397–427.
Raval, Devesh. 2020. "Testing the Production Approach to Markup Estimation." Unpublished.
Rinz, Kevin. 2020. "Labor Market Concentration, Earnings, and Inequality." *Journal of Human Resources* 57: S251–83.
Robinson, Joan. 1933. *The Economics of Imperfect Competition.* New York: Palgrave Macmillan.
Rossi-Hansberg, Esteban, Pierre-Daniel Sarte, and Nicholas Trachter. 2020. "Diverging Trends in National and Local Concentration." *NBER Macroeconomics Annual* 35: 115–50.
Schubert, Gregor, Anna Stansbury, and Bledi Taska. 2021. "Employer Concentration and Outside Options." Unpublished.
Seegmiller, Bryan. 2021. "Valuing Labor Market Power: The Role of Productivity Advantages." Unpublished.
Sokolova, Anna, and Todd Sorensen. 2021. "Monopsony in Labor Markets: A Meta-Analysis." *ILR Review* 74 (1): 27–55.
Staiger, Douglas O., Joanne Spetz, and Ciaran S. Phibbs. 2010. "Is There Monopsony in the Labor Market? Evidence from a Natural Experiment." *Journal of Labor Economics* 28 (2): 211–36.
Starr, Evan P., J.J. Prescott, and Norman D. Bishara. 2021. "Noncompete Agreements in the US Labor Force." *Journal of Law and Economics* 64 (1): 53–84.
Syverson, Chad. 2004a. "Market Structure and Productivity: A Concrete Example." *Journal of Political Economy* 112 (6): 1181–1222.
Syverson, Chad. 2004b. "Product Substitutability and Productivity Dispersion." *Review of Economics and Statistics* 86 (2): 534–50.
Syverson, Chad. 2019. *Macroeconomics and Market Power: Facts, Potential Explanations and Open Questions*. Washington, DC: Brookings Institute.
Traina, James. 2018. "Is Aggregate Market Power Increasing? Production Trends Using Financial Statements." Unpublished.
US Census Bureau. 1976-2014. "Annual Survey of Manufactures."
US Census Bureau. 1977-2012 (quinquennial). "Census of Manufactures."
US Census Bureau. 1976-2014. "Longitudinal Business Database."
Van Reenen, John. 1996. "The Creation and Capture of Rents: Wages and Innovation in a Panel of U.K. Companies." *Quarterly Journal of Economics* 111 (1): 195–226.
Yeh, Chen, Claudia, Macaluso, and Brad Hershbein. 2022. "Monopsony in the US Labor Market." American Economic Association [Publisher], Inter-university Consortium for Political and Social Research [Distributor]. https://doi.org/10.3886/E162581V1.