Labor Market Power*

David Berger Kyle Herkenhoff Simon Mongey

November 9, 2021

Abstract

We develop, estimate, and test a tractable general equilibrium model of oligopsony with differentiated jobs and concentrated labor markets. We estimate key model parameters by matching new evidence on the relationship between firms' local labor market share and their employment and wage responses to state corporate tax changes. The model quantitatively replicates quasi-experimental evidence on imperfect productivity-wage pass-through and strategic wage-setting of dominant employers. Relative to the efficient allocation, welfare losses from labor market power are 7.6 percent, while output is 20.9 percent lower. Lastly, declining local concentration added 4 ppt to labor's share of income between 1977 and 2013.

JEL codes: E2, J2, J42

Keywords: Labor markets, Market structure, Oligopsony, Strategic interaction

*Berger: Duke University. Herkenhoff: Federal Reserve Bank of Minneapolis and University of Minnesota. Mongey: Federal Reserve Bank of Minneapolis and Kenneth C. Griffin Department of Economics, University of Chicago. We thank Costas Arkolakis, Chris Edmond, Oleg Itskohki, Matt Notowidigdo, Richard Rogerson, Jim Schmitz, Chris Tonetti, and Arindrajit Dube for helpful comments. We thank seminar participants at the SED 2018, SITE, Briq (Bonn), FRB Minneapolis, Princeton, Queens, Stanford, UC Berkeley, USC, UT Austin, Rochester, BFI Firms and Workers Workshop, Paris School of Economics, Toulouse School of Economics, EIEF, Barcelona GSE Summer Forum, Stanford, MIT, Boston University, NBER Summer Institute. We thank Chengdai Huang for excellent research assistance. This research was supported by the National Science Foundation (Award No. SES-1824422). Any opinions and conclusions expressed herein are those of the author(s) and do not necessarily represent the views of the U.S. Census Bureau. All results have been reviewed to ensure that no confidential information is disclosed. The views expressed in this study are those of the author and do not necessarily reflect the position of the Federal Reserve Bank of Minneapolis or the Federal Reserve System.

The average local labor market in the U.S. has many firms but employment and wages are concentrated in only a few firms.[1] While the average number of firms is over 100, weighted-average market payroll concentration is 0.11, the same level of concentration one would observe with only 9 equally sized firms.[2] This has led to growing concern that these firms may act strategically and exert "labor market power" over their workers, generating large welfare losses.[3] In this paper we develop a tractable, quantitative, general equilibrium model where jobs are differentiated, local labor markets vary in concentration, and firms behave strategically under an oligopsony equilibrium. These novel features allow the model to quantitatively replicate empirical regularities in the labor literature such as incomplete wage pass-through and strategic competitor wage responses. We use the model to measure the amount of oligopsony power in labor markets and quantify its consequences for output and welfare. The model delivers a structurally consistent formulation of labor market power and a framework for understanding the mechanisms behind potential output and welfare losses.

Our benchmark oligopsony model features two sources of market power. First is classical *monopsony*. From the point of view of each worker, preference heterogeneity implies that jobs are differentiated. Therefore even atomistically small firms face upward sloping labor supply curves, which they internalize (K. Burdett and D.T. Mortensen, 1998; Alan Manning, 2003; Card et al., 2018; Lamadon, Mogstad and Setzler, 2019). Optimal wages are therefore a markdown relative to competitive wages, i.e. the marginal revenue product of labor. Second is *oligopsony*, which is motivated by the level of market concentration that we compute, and is the focus of this paper. Firms are non-atomistic and compete strategically for workers, further internalizing how they expect other employers to respond to their hiring and wage policies. This strategic interaction leads to larger equilibrium markdowns at the most productive firms which generates a second source of welfare loss. Hence, in an oligopsonistic economy, understanding the macroeconomic implications of labor market power requires understanding how markdowns vary across firms. In our model, the markdown is an exact function of the *structural labor supply elasticity* that a firm faces in equilibrium which—via a closed-form—depends on the firm's observable labor market share and parameters that determine how easily labor is reallocated across- (θ) and within- (η) markets.

We estimate the model on U.S. Census data, and derive three main results. First, the framework is quantitatively consistent with documented empirical regularities: incomplete wage pass-through, and competitor wage responses, which is particularly supportive of oligopsony. Qualitatively, two 'monopsony limits' of our model—infinitely many firms in each market, or labor having the same mobility both within and across markets—fail to match these empirical regularities. Second, the model implies substantial output and welfare losses from labor market power. Welfare losses are large, ranging from 6 to 10 percent of lifetime consumption depending on wealth effects, while output losses are even larger

[1]Throughout the paper we define a labor market as the combination of a commuting zone and three-digit industry.

[2]Data is Census LBD for the whole US economy in 2014, see Appendix C for additional details and market level summary statistics. Market payroll concentration is payroll weighted across markets. Appendix Figure A1 plots the distribution of markets and wage payments by concentration. Table A2 provides additional data on employment HHI's. Appendix Table F2 reports 113 firms per market across all industry codes.

[3]For example: José Azar, Ioana Marinescu and Marshall I. Steinbaum (2020), Efraim Benmelech, Nittai Bergman and Hyunseob Kim (2020), David Card, Ana Rute Cardoso, Jörg Heining and Patrick Kline (2018), and Thibaut Lamadon, Magne Mogstad and Bradley Setzler (2019).

ranging from 11 to 31 percent. We derive a representative firm formulation of our economy that delivers equilibrium aggregate prices and quantities and decomposes these losses into two components: (1) a dead-weight loss due to average markdowns, (2) a misallocation effect due to wider markdowns at more productive firms. While the former channel exists in the nested monopsony limits, the latter does not. We find these channels account equally for output and welfare losses. Thus not modeling strategic interaction explicitly would lead a researcher to miss half of the losses due to labor market power. Third, we find that labor market power has not contributed to the declining labor share. Despite the backdrop of stable *national* concentration, we compute for the first time the model-consistent measure of *local* concentration in Census data, and find that it has declined over the last 35 years. Most local labor markets are more competitive than they were in the 1970s.[4]

We prove two theoretical properties of our model that are central to our main applications. First, we show that our model is block recursive, meaning that local labor market equilibria are independent of aggregates. This property allows us to estimate the model efficiently and decompose the macroeconomic implications of labor market power for arbitrary aggregate preferences. Second, we provide a closed-form relationship between labor's share of income and local payroll concentration. Our model-relevant measure of payroll concentration is new to the literature. We use our formula to measure the contribution of changes in local payroll concentration to changes in labor's share of income.

In terms of quantification of the model, we show how strategic interaction complicates the identification of the key parameters. Generically, i.e. away from the two monopsony limits, we show that the Stable Unit Treatment Value Assumption (SUTVA) does not hold, such that exclusion restrictions that are otherwise applicable in monopsonistically competitive models fail. In practice, following a quasi-experiment that yields a shock to labor demand, a researcher can estimate *reduced-form labor supply elasticities* from firm-level employment and wage responses. The literature to date has assumed a nested, special case of our model: firms do not behave strategically, rationalized by infinitely many firms in each labor market.[5] This assumption implies that estimated *reduced-form elasticities* are equal to *structural elasticities*, so one can move directly from empirical analysis to welfare analysis (see Figure 1). In the general case of granular labor markets, there is no closed-form mapping between (observed) reduced-form elasticities and (unobserved) structural elasticities.[6] A model is needed to account for the equilibrium best responses that determine the mapping between underlying structural parameters and the reduced-form elasticities we observe.

Our approach is therefore indirect inference, in which we use U.S. Census Longitudinal Business Database (LBD) micro data to construct reduced form elasticities. Our quasi-experiment used to estimate *reduced-form labor supply elasticities* is motivated by Xavier Giroud and Joshua Rauh (2019), and exploits

[4]In contemporaneous work Kevin Rinz (2018) also uses Census data and shows similar patterns for alternative measures of concentration. These measures are not exactly those that are welfare relevant for the model. Esteban Rossi-Hansberg, Pierre-Daniel Sarte and Nicholas Trachter (2018) use NETS data and find similar patterns in sales and employment concentration.

[5]Papers in the literature that study strategic behavior have been theoretical, which we discuss below.

[6]The finitely many firms case is indeed more general. That is, a 'competitive' monopsony model is indeed a special case of our model. Taking the number of firms in all markets in our model toward infinity smoothly yields the 'competitive' economy in which there is no strategic interaction. We let the data tell us where we are on this spectrum between one and infinitely many firms per market.

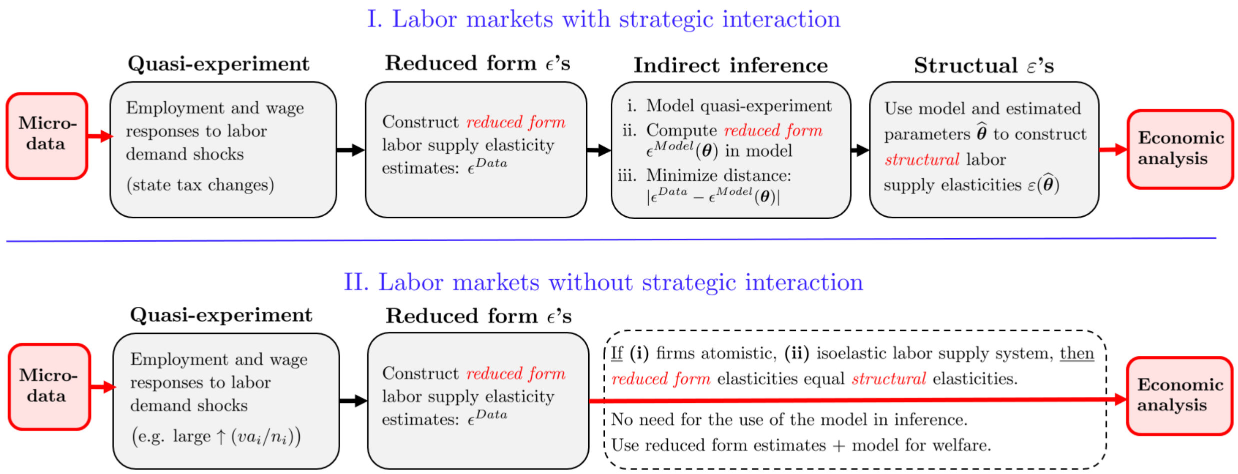

Figure 1: Quantitative strategy

changes in state corporate taxes. We characterize for the first time how firms' employment and wage responses depend on a firm's share of its local labor market. We then simulate tax changes in our model. In the simulated data, the level of reduced form elasticities and their gradient by market share, identify key parameters. The estimated model is then used to compute structural elasticities, markdowns, and conduct welfare counterfactuals.[7]

This departure from the literature contributes three additional results. First, in the data, responses of firms to labor demand shocks vary systematically: firms with smaller market shares have statistically significantly larger reduced-form elasticities than firms with larger market shares, consistent with the prediction of our model. Second, in our particular experiment, *reduced-form elasticities* at small firms are around 2, but welfare-relevant *structural elasticities* are around 10. Filtering the data through the model is necessary to uncover the high labor supply elasticities faced by small firms. Third, we explore bias in more common empirical settings that estimate labor supply elasticities by leveraging instruments for firm labor demand. Here results are different. Even with an ideal instrument for labor demand, *reduced form elasticities* are contaminated by competitors' equilibrium responses and are always less than the underlying *structural elasticities*, often by a large amount. We conclude that a researcher using reduced-form estimates for welfare analysis would infer that firms face flatter labor supply curves and *understate*, the degree of labor market power in the economy.

We further validate the estimated model by replicating two reduced-form experiments that help distinguish empirically between monopsonistic competition and oligopsony in our model. In both cases our model estimates of key elasticities align closely with empirical estimates. First, we replicate the 0.47 pass-through from log value added per worker to log wages in Patrick Kline, Neviana Petkova, Heidi Williams and Owen Zidar (2019), producing 0.50 in our model. Second, we replicate the 0.13 re-

[7] This procedure has a direct counterpart in the estimation of linearized state-space systems in macroeconomics: $A\boldsymbol{X}_t = B\mathbb{E}[\boldsymbol{X}_{t+1}] + C\boldsymbol{X}_{t-1} + D\varepsilon_t$. The *structural model* implies a *reduced-form* VAR representation: $\boldsymbol{X}_{t+1} = H\boldsymbol{X}_t + F\boldsymbol{e}_{t+1}$. The researcher first estimates the reduced-form on the data to obtain *reduced-form* shocks $\{\widehat{e}_t\}_{t=0}^T$. They then simulate *structural shocks* $\{\varepsilon_t\}_{t=0}^T$ in the model and jointly estimate *structural parameters* $\{A, B, C, D\}$ and *structural shocks* $\{\varepsilon_t\}_{t=0}^T$ such that the model implied reduced-form shocks match those obtained from the data.

sponse elasticity of competing hospital's wages to VA hospital wage increases in Douglas O. Staiger, Joanne Spetz and Ciaran S. Phibbs (2010), producing 0.11 in our model. Theoretically, we prove that a monopsonistically competitive version of our economy features a pass-through of one and a competitor response elasticity of zero. These tests provide evidence that oligopsony delivers key empirical regularities in the reduced-form literature.[8]

With our model calibrated to aggregates and local labor markets, we define the *welfare loss* due to labor market power as the consumption subsidy required to make households indifferent between the oligopsonistic economy and the efficient allocation that a planner would choose. Comparing steady states at an aggregate Frisch elasticity of labor supply of 0.50, we measure a welfare loss of 7.6 percent and an output loss of 20.9 percent. Wages and employment would also significantly increase. These results are robust to aggregate preferences being of Jeremy Greenwood, Zvi Hercowitz and Gregory W Huffman (1988, henceforth GHH) or separable types.[9]

To explore the mechanisms underlying these macroeconomic outcomes we provide a novel representative agent counterpart to our economy that decomposes output losses into two components. The first component is an aggregate markdown which reflects pure dead-weight loss from oligopsony power. The second component is an aggregate efficiency loss that reflects misallocation. Productive firms have the most labor market power and widest markdowns. They therefore restrict employment the most. This results in an inefficient under allocation of employment at the most productive firms. Overall, we find that roughly 50 percent of welfare losses are driven by misallocation, 40 percent are due to pure markdowns, and the remainder is due to their interaction.[10] This would not be the case in a monopsonistically competitive version of our economy. The misallocation effect is zero in a monopsonistically competitive version of our economy. Hence, in our economy, strategic interactions and markdown heterogeneity account for roughly half of the losses observed.

A symptom of the misallocation present in the benchmark economy is that the planner's solution has *greater* concentration, employment, and wages. In the oligopsonistic economy, large firms are inefficiently small, so any policy that decentralizes the efficient allocation would reallocate more employment to already large firms. Concentration more than doubles, employment increases by 12 percent and the average wage increases by 43 percent. Importantly, this suggests caution should be exercised in cases where observed changes in concentration are used to make statements about changes in welfare.

We conclude by applying the model to study the relationship between local labor market concentration and the labor share. We find that declining local labor market concentration between 1977 and 2013 *increased* labor's share of income. First, letting our model guide measurement, we show that the distribution of market-level payroll Herfindahls can be used to compute a sufficient statistic for labor's share

[8] Non-homothetic preferences or production technologies may also be able to match the pass-through observation, but we view the competitor best response elasticity as a direct test of our theory.

[9] With more significant wealth effects on labor supply, welfare losses are smaller, but still exceed 5 percent even with a coefficient of relative risk aversion of four. With a higher Frisch elasticity of labor supply, welfare losses are larger. Under an aggregate Frisch of 0.2 (0.8), welfare losses are 5.7 (9.6) percent.

[10] With more significant wealth effects on labor supply, welfare losses due to misallocation increases. With a higher Frisch elasticity of labor supply, welfare losses due to the aggregate markdown increases.

of income, with a relationship that is independent of the aggregate labor supply elasticity and wealth effects.[11] Second, the model implies that these micro measures should be aggregated using market-level payroll weights. We construct this model relevant concentration measure directly from the Census LBD and find it has declined from 0.16 to 0.11 between 1977 and 2013.[12] Ignoring these weights would double the level of concentration and imply a stable trend.[13] We feed our measure into our formula for labor's share of income under the estimated preference parameters (θ, η). We find that declining local labor market concentration would have implied a counterfactual 4 percentage point increase in labor's share of income. Changing labor market concentration is not behind the declining labor share.[14]

We review the literature and then proceed as follows. Sections 1 lays out the model and characterizes the equilibrium. Section 2 provides empirical estimates of the relationship between reduced-form labor supply elasticities and market share, then combine this relationship and our new concentration statistics to parameterize the model. Section 3 validates the model via replication of two empirical studies. Section 4 presents our main welfare measurement exercises. Section 5 applies the model to measure welfare-relevant aggregate concentration and the labor share.

Literature. Our work is related to a growing literature that explores the implications of market power. In the product market, Germán Gutiérrez and Thomas Philippon (2017); David Autor, David Dorn, Lawrence F Katz, Christina Patterson and John Van Reenen (2020) all document an increase in national sales concentration and a fall in the labor share across many industries, while Jan De Loecker, Jan Eeckhout and Gabriel Unger (2020) document an increase in product market power more directly by measuring firm markups. Consistent with our findings, concurrent work by Rossi-Hansberg, Sarte and Trachter (2018) documents declining regional employment concentration, despite rising national concentration. In the labor market, several concurrent studies have documented cross-sectional and time-series patterns of U.S. Herfindahls in employment (Benmelech, Bergman and Kim, 2020; Rinz, 2018; Brad Hershbein, Claudia Macaluso and Chen Yeh, 2020) and vacancies (José Azar, Ioana Marinescu, Marshall Steinbaum and Bledi Taska, 2020; Azar, Marinescu and Steinbaum, 2020). Wyatt J. Brooks, Joseph P. Kaboski, Yao Amber Li and Wei Qian (2019), Hershbein, Macaluso and Yeh (2020), and Mons Chan, Sergio Salgado and Ming Xu (2020) use tools from industrial organization to identify wage markdowns and heterogeneous pass-through rates consistent with the theory in this paper. Our contributions to this literature are (i) a new, model consistent, measure of U.S. labor market concentration, which we use to (ii) quantitatively measure the welfare losses associated with labor market power. In general, the exercises

[11] The market-level wage-bill Herfindahl is the sum of the squared payroll shares of all firms within the labor market

[12] These measures of concentration are equivalent to what would be obtained with 6.25 equally sized firms per market in 1977, and 9.43 equally sized firms per market in 2013.

[13] Our model replicates the distribution and means of both weighted and unweighted Herfindahls in the data. The large difference between weighted and unweighted Herfindahls is due to the fact that 11 percent of markets have one firm, and thus a Herfindahl of 1, yet these markets only comprise 0.18 percent of aggregate payroll. Moreover, the payroll share of concentrated markets is falling, presumably as individuals leave highly concentrated rural markets for less concentrated city markets.

[14] Interestingly, in their recent paper on the dynamics of the labor share, Matthias Kehrig and Nicholas Vincent (2021) find evidence consistent with our results, as employment reallocation is roughly independent of output reallocation (see their Figure III).

in our paper issue a warning against qualitatively mapping changes in concentration into a change in welfare.

Our work is also related to a large literature measuring reduced-form labor supply elasticities of individual firms (Staiger, Spetz and Phibbs, 2010; Douglas A Webber, 2015; Card et al., 2018; Juan Carlos Suárez Serrato and Owen Zidar, 2016; Arindrajit Dube, Jeff Jacobs, Suresh Naidu and Siddharth Suri, 2020). We provide new estimates of reduced-form labor supply elasticities by using regressions motivated by Giroud and Rauh (2019), who find significant effects of state corporate taxes on firm-state employment.[15] Our contributions to this empirical literature are (i) estimates of the share-dependency of reduced-form elasticities that point to large firms having more market power (ii) to demonstrate that if markets have firms that interact strategically, there can be a large disconnect between the reduced-form labor supply elasticities measured by such regressions and the structural elasticities that are relevant for the distribution of labor, and hence welfare. This is a substantive point: the empirical literature cited above typically measures labor supply elasticities that are small. If structural elasticities were equal to these reduced-form elasticities, then labor market power would be extremely high.[16] We describe empirical designs under which (i) reduced-form estimates of labor supply elasticities may be biased downwards relative to structural elasticities, and even then, (ii) that structural elasticities vary systematically with the firm's labor market share. This reconciles the range and level of empirical estimates.

Finally, our work is related to the large literature that models monopsony in labor markets. We depart from benchmark models of monopsony described in Burdett and Mortensen (1998); Manning (2003); Card et al. (2018); Lamadon, Mogstad and Setzler (2019); Kory Kroft, Yao Luo, Magne Mogstad and Bradley Setzler (2020) by explicitly modeling a finite set of employers that compete strategically for workers. We demonstrate that this addition is crucial for identification: strategic interaction and finiteness of firms jointly imply that reduced-form empirical estimates of labor supply elasticities from *any* shock cannot be used to infer the (structural) labor supply elasticities firms face—and hence identify preference parameters—except in the limiting case of monopsonistic competition between infinitesimally sized firms. Additionally, our assumptions allow us to (i) interpret granular measures of concentration, such as Herfindahl indexes, and (ii) accommodate a planning problem that allows us to define an efficient benchmark.

Our main quantitative contribution is to build a general equilibrium model of oligopsony and measure the welfare costs of current levels of U.S. labor market power.[17] Our framework extends the general tools developed in Andrew Atkeson and Ariel Burstein (2008) to the labor market, adding multiple non-

[15]Conceptually, our approach is related to papers that estimate exchange rate pass-through (Mary Amiti, Oleg Itskhoki and Jozef Konings, 2014, 2019). The main difference is that this literature focuses exclusively on prices, whereas we look at both price and quantity responses.

[16]Consider Alan Manning (2011) discussing the widely cited natural experiments of Staiger, Spetz and Phibbs (2010) and others: *"Looking at these studies, one clearly comes away with the impression not that it is hard to find evidence of monopsony power but that the estimates are so enormous to be an embarrassment even for those who believe this is the right approach to labour markets."*

[17]Our work is therefore related to a literature measuring the welfare consequences of misallocation. There the focus has been on the product market (David Rezza Baqaee and Emmanuel Farhi, 2020*b*; Chris Edmond, Virgiliu Midrigan and Daniel Yi Xu, 2018; Jan De Loecker, Jan Eeckhout and Simon Mongey, 2021), and measures misallocation via heterogeneous markups. Our paper measures misallocation from heterogeneous mark-downs.

trivial features: capital, corporate taxes, decreasing returns to scale, and setting the model in general equilibrium. Related contemporaneous work by Gregor Jarosch, Jan Sebastian Nimcsik and Isaac Sorkin (2019) considers non-atomistic firms, but adapts a random search model to construct a search-theoretic measure of labor market power. We view our papers as complementary.

Our model features firm-specific upward sloping labor supply curves. This is supported by numerous recent studies using (quasi-)experimental approaches.[18] Michele Belot, Philipp Kircher and Paul Muller (2017) randomly assign higher wages to observationally equivalent vacancies on an actual job-board and find that higher wage vacancies attract more applicants. Dube et al. (2020) and Stefano Banfi and Benjamin Villena-Roldan (2018) also find job-specific upward sloping labor supply curves in well-identified contexts.[19]

Finally, our quantitative model features strategic complementarity between oligopsonists. Strategic complementarity in labor markets is not new to the theoretical literature. The earliest models used to motivate monopsony power were Joan Robinson (1933) and the spatial economies of Harold Hotelling (1990) and Steven C Salop (1979).[20] Our contribution relative to these stylized single-market models, is a quantitative general equilibrium framework. We incorporate firm heterogeneity, decreasing returns to scale, and general equilibrium across multiple markets, such that the model is rich enough to be estimated on U.S. Census data. Moreover, by modeling a finite set of employers, our model may be used in the future to understand the wage and welfare effects of minimum wages, mergers, firm exit, and other shocks that interact with local labor market competition. Recent work by Miren Azkarate-Askasua and Miguel Zerecero (2020) and Gaelan MacKenzie (2019) also estimate models with strategic interactions using French and Indian data, respectively. Our contribution is to develop a quantitative general equilibrium framework and develop a methodology to consistently estimate the underlying preference parameters governing oligopsony.

1 Model

1.1 Environment

Agents. The economy consists of a representative household and a continuum of firms. The household consists of a unit measure of atomistic, homogeneous workers each with one unit of labor supply. Firms are heterogeneous in two dimensions. First, firms inhabit a continuum of local labor markets $j \in [0,1]$, each with an exogenous and finite number of firms indexed $i \in \{1, 2, \ldots, m_j\}$. Second, firms' productivities $z_{ijt} \in (0, \infty)$ are drawn from a location invariant distribution $F(z)$. The *only ex-ante difference* between markets is the number of firms $m_j \in \{1, \ldots, \infty\}$. Time subscripts are necessary for the household capital accumulation decision, but productivity and number of firms are constant at the firm- and market-level,

[18]See Orley C Ashenfelter, Henry Farber and Michael R Ransom (2010) for a summary of prior papers.

[19]We are unaware of experimental evidence regarding the market-share dependence of the elasticity of labor supply.

[20]William M Boal and Michael R Ransom (1997) and Venkataraman Bhaskar, Alan Manning and Ted To (2002) provide excellent summaries of strategic complementarity in spatial models of the labor market.

respectively.[21]

Goods and technology. The continuum of firms produce tradeable goods that are perfect substitutes, and so trade in a perfectly competitive national market at a price P_t that we normalize to one. Firms operate a *value-added* production function that uses inputs of capital k_{ijt} and labor n_{ijt}.[22] A firm produces y_{ijt} units of net-output (value-added) according to the production function:

$$y_{ijt} = z_{ijt}\left(k_{ijt}^{1-\gamma} n_{ijt}^{\gamma}\right)^{\alpha} \quad , \quad \gamma \in (0,1) \quad , \quad \alpha > 0.$$

The degree of returns to scale α is unrestricted and later estimated. The household uses these goods for consumption and investment. Investment augments the capital stock K_t, which is rented to firms in a competitive market at price R_t and depreciates at rate δ. To the best of our knowledge this is the first paper to model imperfect competition, either in input or output markets, with finitely many firms and decreasing returns to scale in general equilibrium. To model imperfect competition we extend tools developed in the trade literature (Atkeson and Burstein, 2008).

1.2 Household

Preferences and problem. The household chooses the measure of workers to supply to each firm n_{ijt}, investment in next period capital K_{t+1}, and consumption of each good c_{ijt} to maximize their net present value of utility. Given an initial capital stock K_0, the household solves

$$\mathcal{U}_0 = \max_{\left\{n_{ijt}, c_{ijt}, K_{t+1}\right\}_{t=0}^{\infty}} \sum_{t=0}^{\infty} \beta^t U\left(\boldsymbol{C}_t, \boldsymbol{N}_t\right) \tag{1}$$

where the aggregate consumption and labor supply indexes are given by:

$$\boldsymbol{C}_t := \int_0^1 \left[c_{1jt} + \cdots + c_{m_j jt}\right] dj \ , \ \boldsymbol{N}_t := \left[\int_0^1 \boldsymbol{n}_{jt}^{\frac{\theta+1}{\theta}} \, dj\right]^{\frac{\theta}{\theta+1}} \ , \ \boldsymbol{n}_{jt} := \left[n_{1jt}^{\frac{\eta+1}{\eta}} + \cdots + n_{m_j jt}^{\frac{\eta+1}{\eta}}\right]^{\frac{\eta}{\eta+1}} \ , \ \eta > \theta > 0$$

and maximization is subject to the household's budget constraint in each period:

$$\boldsymbol{C}_t + \left[K_{t+1} - (1-\delta)K_t\right] = \int_0^1 \left[w_{1jt} n_{1jt} + \cdots + w_{m_j jt} n_{m_j jt}\right] dj + R_t K_t + \Pi_t. \tag{2}$$

Firm profits, Π_t, are rebated lump sum to the household. The function U is twice continuously differentiable with standard properties.[23] The consumption index captures perfect substitutability of consumption goods, such that our assumption of a single market price $P_t = 1$ is valid.[24]

[21] Earlier drafts of this paper included transition dynamics, yielding similar results to our steady state analysis. The model's block recursivity make transition dynamics tractable. See David W. Berger, Kyle F. Herkenhoff and Simon Mongey (2021*a*).

[22] Since aggregating firm-level value-added yields aggregate output (GDP), we abuse terminology and refer to the output of this production function interchangeably in terms of goods and value-added. We carefully distinguish the two when comparing our results to empirical studies.

[23] Properties: $U_C > 0$, $U_{CC} < 0$, $U_N < 0$, $U_{NN} > 0$, $\lim_{C\to 0} U_C = -\lim_{N\to\infty} U_N = \infty$, $\lim_{C\to\infty} = -\lim_{N\to 0} U_N = 0$.

[24] Observe that since we are solving the model with decreasing returns to scale in production, we are arbitrarily able to introduce monopolistic competition in the national market for goods. Let $C_t = [\int \sum_{i\in j} c_{ijt}^{(\sigma-1)/\sigma} dj]^{\sigma/(\sigma-1)}$, then given household's optimal demand schedules, a firm would optimize a decreasing returns to scale *revenue function* as opposed to the decreasing returns to scale *production function* used here. Firms would charge identical time-invariant markups, and profits due to market power in the product market would be rebated to the household. To keep our analysis clean, we ignore this case.

Notation. Aggregate variables are denoted in upper-case, and firm- and market-level in lower-case. Bold fonts are used for indexes, which are book-keeping devices, not directly observable in the raw data, but can be constructed from observables. For example, the disutility of labor supply $\boldsymbol{N}_t$ does not correspond to any aggregates reported by the Bureau of Labor Statistics. However, given parameters, $\boldsymbol{N}_t$ can be constructed from the universe of firm-level employment $\{n_{ijt}\}$. We denote aggregate labor computed by adding workers as unbolded: $N_t = \int \sum_i n_{ijt}\, dj$.

Optimality conditions. The first order necessary conditions of the household problem describe the supply of labor and capital:

$$-\frac{U_N\left(\boldsymbol{C}_t, \boldsymbol{N}_t\right)}{U_C\left(\boldsymbol{C}_t, \boldsymbol{N}_t\right)} \frac{\partial \boldsymbol{N}_t}{\partial \boldsymbol{n}_{jt}} \frac{\partial \boldsymbol{n}_{jt}}{\partial n_{ijt}} = w_{ijt} \quad , \quad U_C\left(\boldsymbol{C}_t, \boldsymbol{N}_t\right) = \beta U_C\left(\boldsymbol{C}_{t+1}, \boldsymbol{N}_{t+1}\right)\left[R_t + (1-\delta)\right] \tag{3}$$

Labor supply. Under the assumed structure of preferences, we can express the set of labor supply conditions across all firms more economically as follows:

$$\underbrace{-\frac{U_N\left(\boldsymbol{C}_t, \boldsymbol{N}_t\right)}{U_C\left(\boldsymbol{C}_t, \boldsymbol{N}_t\right)} = \boldsymbol{W}_t}_{\text{Aggregate labor supply}} \quad \text{and} \quad \underbrace{n_{ijt} = \left(\frac{w_{ijt}}{\boldsymbol{w}_{jt}}\right)^{\eta} \left(\frac{\boldsymbol{w}_{jt}}{\boldsymbol{W}_t}\right)^{\theta} \boldsymbol{N}_t}_{\text{Firm labor supply for all } i=1,\ldots,m_j,\, j\in[0,1].} \quad \leftrightarrow \quad \underbrace{w_{ijt} = \left(\frac{n_{ijt}}{\boldsymbol{n}_{jt}}\right)^{\frac{1}{\eta}} \left(\frac{\boldsymbol{n}_{jt}}{\boldsymbol{N}_t}\right)^{\frac{1}{\theta}} \boldsymbol{W}_t.}_{\text{Inverse labor supply curve}} \tag{4}$$

Given aggregate labor supply, the firm labor supply curve includes two book-keeping terms: the *market wage index* $\boldsymbol{w}_{jt}$ and *aggregate wage index* $\boldsymbol{W}_t$. These are defined as the numbers that satisfy

$$\boldsymbol{w}_{jt}\boldsymbol{n}_{jt} := \sum_{i\in j} w_{ijt} n_{ijt} \quad , \quad \boldsymbol{W}_t \boldsymbol{N}_t := \int_0^1 \boldsymbol{w}_{jt}\boldsymbol{n}_{jt}\, dj.$$

Together with optimality conditions (4) these definitions imply

$$\boldsymbol{w}_{jt} = \left[\sum_{i\in j} w_{ijt}^{1+\eta}\right]^{\frac{1}{1+\eta}} \quad , \quad \boldsymbol{W}_t = \left[\int_0^1 \boldsymbol{w}_{jt}^{1+\theta} dj\right]^{\frac{1}{1+\theta}}. \tag{5}$$

Since labor market competition is Cournot, firms choose quantities taking their inverse labor supply curve (4) into account. For full derivations see Appendix E.1.

Explicit Microfoundation. In Appendix B, we show that the supply system described by equations (4) and (5) can be obtained in an environment with heterogeneous workers making independent decisions, providing an exact map between η and θ and the distribution of relative net costs to individuals of moving between and across markets.[25] The micro-foundation makes clear that workers are not confined to particular markets. The limitation that markets impose is on the boundary of the strategic behavior of firms. Within markets firms are strategic, but with respect to firms in the continuum of other markets, firms are price takers.

[25] Recent (non-nested) logit formulations of individual decisions have also been used to model the supply of labor to a firm in competitive markets (Card et al., 2018; Katarina Borovickova and Robert Shimer, 2017). Our contribution is to adapt results in the discrete choice literature to demonstrate equivalence with our 'nested-CES' specification, and to set the problem in oligopsonistic markets. In particular, we adapt arguments from the product market case due to Frank Verboven (1996). That paper the establishes the equivalence of nested-logit and nested-CES, extending the results of Simon P. Anderson, Andre De Palma and Jacques-Francois Thisse (1987) which establishes an equivalence between single sector CES and single sector logit.

Elasticities. The firm labor supply curve is upward sloping and features two elasticities of substitution $\eta > 0$ and $\theta > 0$. These jointly affect the labor market power of firms. Both across and within markets, the lower the degree of substitutability, the greater the market power of firms. Across-market substitutability θ stands in for mobility costs across markets, which are often estimated to be significant (John Kennan and James R Walker, 2011). As such costs increase ($\theta \to 0$), the household minimizes labor disutility $\boldsymbol{N}_t$ by choosing an equal division of workers across markets: $\boldsymbol{n}_{jt} = \boldsymbol{n}_{j't}, \forall j, j' \in [0,1]$. This imparts the largest degree of local labor market power as market-by-market, market-level employment becomes perfectly inelastic and unresponsive to across-market wage differences. As substitutability approaches infinity, the representative household optimally sends all workers to the market with the highest wage, eroding market power of firms in competing markets.

Within-market substitutability η stands in for within-market, across-firm mobility costs such as the job search process (Burdett and Mortensen, 1998), some degree of non-generality of accumulated human capital (Gary S Becker, 1962), or preference heterogeneity in the form of worker-firm specific amenities or commuting costs (Robinson, 1933). As these costs increase ($\eta \to 0$), the household minimizes within-market disutility $\boldsymbol{n}_{jt}$ by choosing an equal division of workers across firms: $n_{ijt} = n_{i'jt}, \forall i, i' \in \{1, 2, ...m_j\}$. This generates the largest degree of monopsony power to firms within a market. Regardless of its wage, firm-ij will employ the same number of workers, allowing it to pay less while maintaining its workforce. As substitutability increases, competition tightens as workers are reallocated toward firms with higher wages.

Regardless of θ, in the limit as $\eta \to \infty$, local labor markets tend to perfect competition. In this limit, marginal revenue products are equalized across firms at a single market wage $w_{ij} = \boldsymbol{w}_j$. This is possible with productivity heterogeneity due to decreasing returns as in Hugo A. Hopenhayn (1992). A model without decreasing returns would mistakenly infer labor market power from the fact that there is productivity heterogeneity and many firms operate in each market.

1.3 Firms

In order to maximize profits, firms choose how much capital to rent, k_{ijt}, and the number of workers to hire n_{ijt}. Infinitesimal with respect to the macroeconomy, firms take the aggregate wage $\boldsymbol{W}_t$ and labor supply $\boldsymbol{N}_t$ as given. Since the equilibrium concept is Cournot, they also take as given their competitors' employment decisions, which we denote n^*_{-ijt}.

The firm maximizes profits:

$$\pi_{ijt} = \max_{n_{ijt}, k_{ijt}} \underbrace{z_{ijt}\left(k_{ijt}^{1-\gamma} n_{ijt}^{\gamma}\right)^{\alpha}}_{\text{Value added: } y_{ijt}} - R_t k_{ijt} - w\left(n_{ijt}, n^*_{-ijt}, \boldsymbol{N}_t, \boldsymbol{W}_t\right) n_{ijt}. \tag{6}$$

$$s.t. \quad w\left(n_{ijt}, n^*_{-ijt}, \boldsymbol{N}_t, \boldsymbol{W}_t\right) = \left(\frac{n_{ijt}}{\boldsymbol{n}(n_{ijt}, n^*_{-ijt})}\right)^{\frac{1}{\eta}} \left(\frac{\boldsymbol{n}(n_{ijt}, n^*_{-ijt})}{\boldsymbol{N}_t}\right)^{\frac{1}{\theta}} \boldsymbol{W}_t \quad , \quad \boldsymbol{n}(n_{ijt}, n^*_{-ijt}) = \left[n_{ijt}^{\frac{\eta+1}{\eta}} + \sum_{k \neq i} n^{*\,\frac{\eta+1}{\eta}}_{kjt}\right]^{\frac{\eta}{\eta+1}}$$

The first order necessary conditions of the firm problem describe its demand for capital and labor:

$$R_t = \alpha(1-\gamma)\frac{y_{ijt}}{k_{ijt}} \quad , \quad \underbrace{w_{ijt} + \left.\frac{\partial w_{ijt}}{\partial n_{ijt}}\right|_{n^*_{-ijt}} n_{ijt}}_{\text{Marginal cost: } mc_{ij}} = \alpha\gamma\frac{y_{ijt}}{n_{ijt}}.$$

The firm has a standard competitive demand for capital, but since the firm has market power in the labor market, its marginal cost of labor accounts for both the wage and the additional cost associated with increasing wages. This requires an equilibrium marginal revenue product of labor that exceeds the wage alone. The standard re-arrangement of the labor demand condition yields a Lerner condition for the wage as a markdown $\mu_{ijt} \leq 1$ on the marginal product of labor:

$$w_{ijt} = \mu_{ijt}\alpha\gamma\frac{y_{ijt}}{n_{ijt}} \quad , \quad \mu_{ijt} = \frac{\varepsilon_{ijt}}{\varepsilon_{ijt}+1} \quad , \quad \varepsilon_{ijt} := \left[\left.\frac{\partial \log w_{ijt}}{\partial \log n_{ijt}}\right|_{n^*_{-ijt}}\right]^{-1}. \tag{7}$$

Under our specification of preferences, the elasticity and markdown have closed-form expressions that depend only on firms' payroll share s_{ijt} in the market, with larger firms having wider markdowns:

$$\varepsilon(s_{ijt}) = \left[\frac{1}{\eta} + \left(\frac{1}{\theta}-\frac{1}{\eta}\right)\left.\frac{\partial \log \boldsymbol{n}_{jt}}{\partial \log n_{ijt}}\right|_{n^*_{-ijt}}\right]^{-1} = \left[\frac{1}{\eta} + \left(\frac{1}{\theta}-\frac{1}{\eta}\right)s_{ijt}\right]^{-1} \quad , \quad s_{ijt} := \frac{w_{ijt}n_{ijt}}{\sum_{i=1}^{m_j} w_{ijt}n_{ijt}} = \frac{w_{ijt}n_{ijt}}{\boldsymbol{w}_{jt}\boldsymbol{n}_{jt}}.$$

When a firm is infinitesimal changes in its employment do not move market employment $\boldsymbol{n}_{jt}$ and hence its labor supply elasticity is simply η, reflecting within market substitution. When a firm is large, its effect on $\boldsymbol{n}_{jt}$ implies it takes into account the lower elasticity of substitution across markets θ. We characterize the solution of the economy in three steps: partial equilibrium, market equilibrium, and general equilibrium.

1.4 Characterization - Partial equilibrium

It will be useful to substitute the firms' capital demand condition into its profits (6), which gives:

$$\pi_{ijt} = \max_{n_{ijt}} \ \widetilde{z}_{ijt} n_{ijt}^{\widetilde{\alpha}} - w_{ijt}n_{ijt} \quad , \quad \text{subject to the inverse labor supply curve (4),}$$

where we introduce the auxiliary parameters $\{\widetilde{\alpha}, \widetilde{z}_{ijt}\}$:

$$\widetilde{\alpha} := \frac{\gamma\alpha}{1-(1-\gamma)\alpha} \quad , \quad \widetilde{z}_{ijt} := \left[1-(1-\gamma)\alpha\right]\left(\frac{(1-\gamma)\alpha}{R_t}\right)^{\frac{(1-\gamma)\alpha}{1-(1-\gamma)\alpha}} z_{ijt}^{\frac{1}{1-(1-\gamma)\alpha}}.$$

We can then express the *markdown* ($\mu_{ijt} \in (0,1)$), marginal and average product of labor as:[26]

$$w_{ijt} = \mu\left(s_{ijt}\right)\, mrpl_{ijt} \quad , \quad mrpl_{ijt} = \widetilde{\alpha}\widetilde{z}_{ijt}n_{ijt}^{\widetilde{\alpha}-1} \quad , \quad arpl_{ijt} = \frac{1}{\widetilde{\alpha}}mrpl_{ijt} \tag{8}$$

with the same formulas as above determining the markdown.

Figure 2 characterizes firm optimality. Decreasing returns implies a downward sloping marginal

[26]Here we have abused description slightly since we are using a value-added production function and maximized out optimal capital, so this is really the marginal "revenue net of capital and intermediate input expense" product of labor.

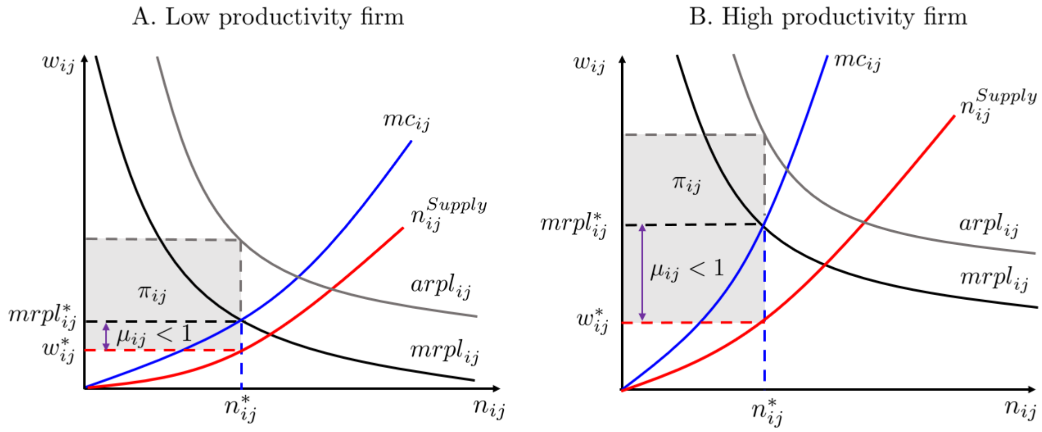

Figure 2: Firm level optimality

revenue product of labor strictly below the average revenue product. Firms internalize their upward sloping labor supply curve, so their marginal cost of labor is also upward sloping and lies strictly above labor supply which describes the average cost of labor. At the margin, a unit of labor costs more than just the higher wage paid to the marginal worker, since the firm must increase wages paid to all workers. As such, choosing n_{ijt} so that labor's marginal revenue product equals its marginal cost necessarily implies a *markdown* of the wage relative to marginal revenue product. The firm then earns profits of $arpl_{ij} - w_{ij} = (arpl_{ij} - mrpl_{ij}) + (mrpl_{ij} - w_{ij})$ from each worker, with a contribution due to the gap between average and marginal revenue products, and a gap due to the markdown.

These markdowns constitute our measure of *firm level labor market power*, and depend on firm characteristics. As we have established, in the Cournot Nash equilibrium, they are determined by the equilibrium (inverse) labor supply elasticity faced by the firm $(1/\varepsilon_{ijt})$ *at the equilibrium allocation*. This depends on a firm's own (observable) market share as well as the degree of within-market (η) and across-market (θ) labor substitutability. This can be seen by returning to Figure 2. Panel A describes the equilibrium outcomes for a low productivity firm. Relative to the high productivity firm in panel B, the low productivity firm has a lower $mrpl_{ij}$ for any n_{ij}. In equilibrium, it has both lower wages w_{ij}^*, and lower employment n_{ij}^*, so its share of wage payments s_{ij}^*, is smaller. With a smaller share of the labor market wage payments, its partial equilibrium elasticity of labor supply is higher, and its inverse labor supply curve is flatter. A flatter inverse supply curve yields a narrower markdown at its optimal labor demand, n_{ij}^*. The larger firm faces an endogenously steeper supply curve and hires more workers at higher wages but at a wider markdown. A key property of this equilibrium is that a larger share of employment is at wide markdown firms.

1.5 Characterization - Market equilibrium

Given firm optimality, we establish properties of the market equilibrium and provide an example which illustrates strategic interactions within the market.

Proposition 1.1. Block recursivity. *In each market $j \in [0,1]$, the equilibrium market shares $s_{1jt}, \ldots, s_{m_j jt}$ satisfy the following m_j equations:*

$$s_{ijt} = \frac{\left[\mu(s_{ijt})^{1-(1-\gamma)\alpha} z_{ijt}\right]^{\frac{\eta+1}{(1-\alpha)(\eta+1)+\alpha\gamma}}}{\sum_{k=1}^{m_j}\left[\mu(s_{kjt})^{1-(1-\gamma)\alpha} z_{kjt}\right]^{\frac{\eta+1}{(1-\alpha)(\eta+1)+\alpha\gamma}}}\ ,\ \mu(s_{ijt}) = \frac{\varepsilon(s_{ijt})}{\varepsilon(s_{ijt})+1}\ ,\ \varepsilon(s_{ijt}) = \left[s_{ijt}\theta^{-1} + (1-s_{ijt})\eta^{-1}\right]^{-1}\ , \forall i = 1, \ldots, m_j \tag{9}$$

This system is independent of aggregate variables. Therefore the joint distribution $\{s_{ijt}, \mu_{ijt}, z_{ijt}\}_{\forall ij}$ is determined under market equilibrium. The labor share at the market level and market payroll concentration are given by the following, and hence independent of aggregates:

$$ls_j = \frac{\sum_{i\in j} w_{ij} n_{ij}}{\sum_{i\in j} y_{ij}} = \Big[\sum_{i\in j} s_{ij} ls_{ij}^{-1}\Big]^{-1} = \alpha\gamma\Big[\sum_{i\in j} s_{ij}\mu_{ij}^{-1}\Big]^{-1} \quad , \quad hhi_j = \sum_{i\in j} s_{ij}^2.$$

Proposition (1.1) establishes that the equilibrium of the model is *block recursive* in that the market equilibrium can be solved without knowledge of aggregate variables. For the proof see Appendix E.3. This has three significant implications. First, solving the Nash equilibrium in a large J number of markets is computationally expensive. Proposition (1.1) says that this need only be done once. Second, the aggregate economy can be arbitrarily rich, and feature transition dynamics that do not require re-solving the J market equilibria. Third, if it can be shown that an aggregate moment of the economy only depends on the joint distribution of markdowns and productivity, then we know that such moments are robust to alternative assumptions on preferences and capital accumulation. Below we will use only these types of moments in our calibration, so that our calibration is robust to assumptions on preferences.

The logic underlying the proof of this proposition is that we can consider the equilibrium for the firm as a recursive set of equations that determine the marginal revenue product of labor:

$$mrpl_{ijt} = \widetilde{\alpha}\widetilde{z}_{ijt} n_{ijt}^{\widetilde{\alpha}-1} \quad , \quad n_{ijt} = \left(\frac{w_{ijt}}{\boldsymbol{w}_{jt}}\right)^{\eta} \boldsymbol{n}_{jt} \quad , \quad w_{ijt} = \mu(s_{ijt}) mrpl_{ijt}.$$

This system implies a multiplicative relationship between $mrpl_{ij}$ and the factors common to all firms in the market: $\boldsymbol{w}_{jt}$, $\boldsymbol{n}_{jt}$. Since payroll shares can be expressed in terms of relative wages $s_{ijt} = (w_{ijt}/\boldsymbol{w}_{jt})^{(1+\eta)}$, the homotheticity of $\boldsymbol{w}_{jt}$ implies that these common factors drop out. For a full proof see Appendix E.

Decreasing returns. The expression for equilibrium payroll shares in Proposition 1.1 is new, and extends such expressions in constant returns oligopoly models to the case of oligopsony, multiple inputs, and decreasing returns. It also provides a novel link between returns to scale and concentration. Consider starting with $\alpha < 1$ and $\gamma = 1$, such that labor is the sole input to production. Now consider the comparative static of increasing α to $\alpha' \in (\alpha, 1]$. With less decreasing returns, more productive firms become larger, accrue a larger labor market share, and pay wider markdowns relative to marginal products. This increases the dispersion in market shares and markdowns in the market, reduces the labor share, and increases concentration.

Example. To show how strategic interaction shapes the market equilibrium, Figure 3 plots examples of the equilibrium shares, markdowns, wages, and employment in three markets. The first market has a

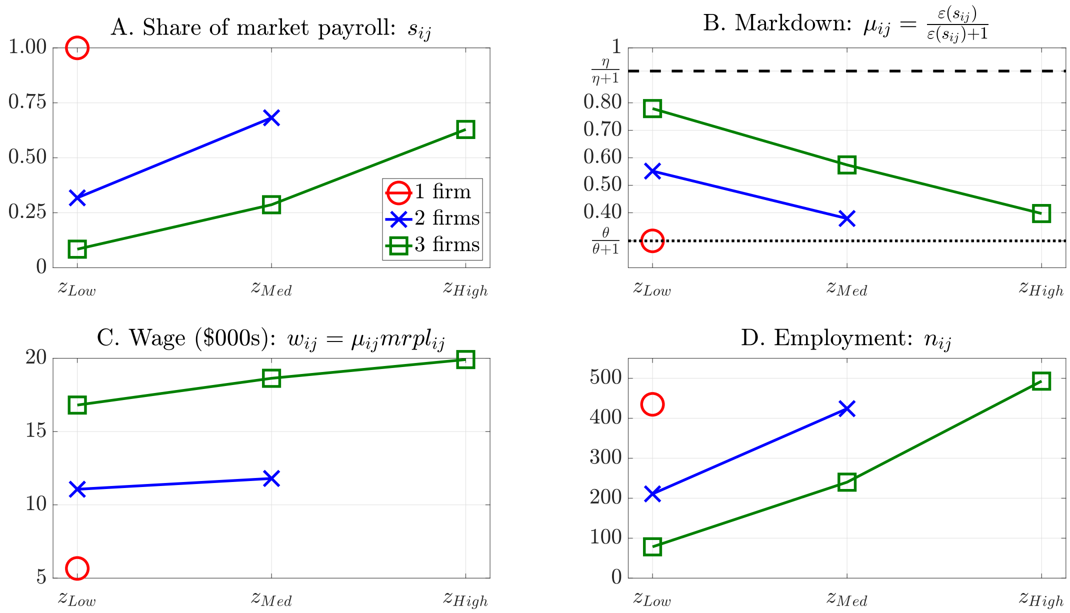

Figure 3: Oligopsonistic market equilibrium in three labor markets

Notes: Figure constructed from model under estimated parameters (Table 3). Low, medium and high productivities of the firms correspond to the 10^{th}, 50^{th} and 90^{th} percentiles of the productivity distribution.

single low productivity firm (red), the second adds a firm with median productivity (blue), the third an additional high productivity firm (green).[27]

Consider the market with a single firm (red). By construction, the wage bill share is one (Panel A). Panel B shows that the markdown on the marginal product of labor is 30 percent, which is equal to $\theta/(\theta+1)$ since the firm faces the lower bound on labor supply elasticities, $\varepsilon(1) = \theta$. Panel C shows that wages are low due to low productivity *and* a wide markdown. Despite this, the relatively inelastic labor supply across markets means the firm still employs many workers (panel D).

Consider the addition of a firm with higher productivity, a duopsony (blue). The low-productivity firm's labor market share drops to 32 percent, the more productive firm employs the majority of the market, and market employment is higher. As its share falls, the low-productivity firm's markdown narrows to 55 percent, as more competition increases their equilibrium labor supply elasticity toward η. Panel C shows that with no change to its productivity, but with a narrower markdown, the less productive firm's wage increases. Despite this wage increase, the higher wage at its new competitor bids away labor, causing the low productivity firm's employment to fall. Adding another firm (green), the markdown at the low- and mid-productivity firms decline. The largest firm has the widest markdown (Panel B), but pays more (Panel C) and employs more workers (Panel D).

Figure A3 replicates this exercise with three firms but varying decreasing returns α. Consistent with our above description, higher α generates more concentration and wider markdowns at the leading firm.

Strategic interaction is not an assumption, it's an outcome of the environment, and leads to a negative

[27]Figure 3 is constructed from our benchmark calibration of the model (Section 2).

covariance between markdowns and productivity—visible along the green line in Panel B. In equilibrium, strategic interaction occurs by definition of the Nash equilibrium concept when there is local labor market power ($\eta > \theta$) and finitely many firms. Under a monopsonistically competitive special case of our model, the green line would be flat, as firms all pay identical markdowns. We now make precise how this negative covariance distorts the general equilibrium of the economy.

1.6 General equilibrium

Given equilibria in each market of the economy, which determines $\{\mu_{ijt}, z_{ijt}\}_{\forall ij}$, we state our main proposition characterizing the general equilibrium of the economy. For the proof see Appendix E.4.

Proposition 1.2. General equilibrium. *The general equilibrium of the model can be characterized in the following three steps:*

1. *Using the market equilibria $\{\mu_{ijt}, z_{ijt}\}_{i=1}^{m_j}$ from all $j \in [0,1]$ markets in the economy, define the following indexes:*

$$
\begin{aligned}
\textit{Productivity}: \quad \widetilde{\boldsymbol{Z}} &= \left[\int_0^1 \widetilde{\boldsymbol{z}}_j^{\frac{1+\theta}{1+\theta(1-\widetilde{\alpha})}} dj\right]^{\frac{1+\theta(1-\widetilde{\alpha})}{1+\theta}} \quad , \quad \widetilde{\boldsymbol{z}}_j = \left[\sum_{i=1}^{m_j} \widetilde{z}_{ij}^{\frac{1+\eta}{1+\eta(1-\widetilde{\alpha})}}\right]^{\frac{1+\eta(1-\widetilde{\alpha})}{1+\eta}} \\
\textit{Markdown}: \quad \boldsymbol{\mu} &= \left[\int_0^1 \left(\frac{\widetilde{\boldsymbol{z}}_j}{\widetilde{\boldsymbol{Z}}}\right)^{\frac{1+\theta}{1+\theta(1-\widetilde{\alpha})}} \boldsymbol{\mu}_j^{\frac{1+\theta}{1+\theta(1-\widetilde{\alpha})}} dj\right]^{\frac{1+\theta(1-\widetilde{\alpha})}{1+\theta}} \quad , \quad \boldsymbol{\mu}_j = \left[\sum_{i=1}^{m_j} \left(\frac{\widetilde{z}_{ij}}{\widetilde{\boldsymbol{z}}_j}\right)^{\frac{1+\eta}{1+\eta(1-\widetilde{\alpha})}} \mu_{ij}^{\frac{1+\eta}{1+\eta(1-\widetilde{\alpha})}}\right]^{\frac{1+\eta(1-\widetilde{\alpha})}{1+\eta}} \\
\textit{Misallocation}: \quad \boldsymbol{\Omega} &= \int_0^1 \left(\frac{\widetilde{\boldsymbol{z}}_j}{\widetilde{\boldsymbol{Z}}}\right)^{\frac{1+\theta}{1+\theta(1-\widetilde{\alpha})}} \left(\frac{\boldsymbol{\mu}_j}{\boldsymbol{\mu}}\right)^{\frac{\widetilde{\alpha}\theta}{1+\theta(1-\widetilde{\alpha})}} \boldsymbol{\omega}_j \, dj \quad , \quad \boldsymbol{\omega}_j = \sum_{i=1}^{m_j} \left(\frac{\widetilde{z}_{ij}}{\widetilde{\boldsymbol{z}}_j}\right)^{\frac{1+\eta}{1+\eta(1-\widetilde{\alpha})}} \left(\frac{\mu_{ij}}{\boldsymbol{\mu}_j}\right)^{\frac{\eta\widetilde{\alpha}}{1+\eta(1-\widetilde{\alpha})}}
\end{aligned}
$$

2. *In steady-state the four aggregate quantities $\boldsymbol{Y}$, $\boldsymbol{N}$, $\boldsymbol{C}$, K and two prices $\boldsymbol{W}$, R are then determined by:*

$$
\begin{aligned}
\textit{Output and resource constraint:} \quad \boldsymbol{Y} &= \boldsymbol{\Omega}^{1-(1-\gamma)\alpha} \boldsymbol{Z} \left(K^{1-\gamma} \boldsymbol{N}^{\gamma}\right)^{\alpha} \quad , \quad \boldsymbol{C} = \boldsymbol{Y} - \delta \boldsymbol{K} \\
\textit{Labor and capital demand:} \quad \boldsymbol{W} &= \gamma\alpha \left(\frac{\boldsymbol{\mu}}{\boldsymbol{\Omega}}\right) \frac{\boldsymbol{Y}}{\boldsymbol{N}} \quad , \quad R = (1-\gamma)\alpha \frac{\boldsymbol{Y}}{K} \\
\textit{Labor and capital supply:} \quad \boldsymbol{W} &= -\frac{U_N(\boldsymbol{C}, \boldsymbol{N})}{U_C(\boldsymbol{C}, \boldsymbol{N})} \quad , \quad 1 = \beta\left[R + (1-\delta)\right]
\end{aligned}
$$

where aggregate productivity $\boldsymbol{Z}$ satisfies [28]

$$
\boldsymbol{Z} = \left[\frac{R}{(1-\gamma)\,\alpha}\right]^{(1-\gamma)\alpha} \left[\frac{\widetilde{\boldsymbol{Z}}}{1-(1-\gamma)\,\alpha}\right]^{1-(1-\gamma)\alpha}
$$

3. *Given aggregate quantities and prices, firm level variables can be obtained as follows. First, equating market labor demand and market labor supply determines $\boldsymbol{w}_j$ and $\boldsymbol{n}_j$. Second, equating firm labor demand and firm labor*

[28] Note that we could directly compute productivity $\boldsymbol{Z}$ using only primitives: $\boldsymbol{z}_j := \left[\sum_{i\in j} z_{ij}^{\frac{1+\eta}{1-(1-\gamma)\alpha+\eta(1-\alpha)}}\right]^{\frac{1-(1-\gamma)\alpha+\eta(1-\alpha)}{1+\eta}}$ and $\boldsymbol{Z} := \left[\int z_j^{\frac{1+\theta}{1-(1-\gamma)\alpha+\theta(1-\alpha)}} dj\right]^{\frac{1-(1-\gamma)\alpha+\theta(1-\alpha)}{1+\theta}}$. Using these as primitives leads to long exponents on the $\boldsymbol{\mu}_j$, $\boldsymbol{\mu}$, $\boldsymbol{\omega}_j$, and $\boldsymbol{\Omega}$ terms, hence we state the proposition in terms of effective productivities after the firms' optimal capital choice.

supply determines w_{ij} and n_{ij}:

$$\boldsymbol{w}_j = \underbrace{\boldsymbol{\mu}_j \widetilde{\alpha} \widetilde{z}_j \boldsymbol{n}_j^{\widetilde{\alpha}-1}}_{\text{Labor demand}} = \underbrace{\left(\frac{\boldsymbol{n}_j}{\boldsymbol{N}}\right)^{1/\theta} \boldsymbol{W}}_{\text{Labor supply}} \quad , \quad w_{ij} = \underbrace{\mu_{ij} \widetilde{\alpha} \widetilde{z}_{ij} n_{ij}^{\widetilde{\alpha}-1}}_{\text{Labor demand}} = \underbrace{\left(\frac{n_{ij}}{\boldsymbol{n}_j}\right)^{1/\eta} \boldsymbol{w}_j}_{\text{Labor supply}} .$$

An alternative, intuitive, representation of the aggregate equations can be obtained using the 'tilde' objects introduced previously, giving four equations determining consumption, output, labor and the wage:

$$\boldsymbol{W} = \underbrace{-\frac{U_N(\boldsymbol{C}, \boldsymbol{N})}{U_C(\boldsymbol{C}, \boldsymbol{N})}}_{\text{Labor supply}} = \underbrace{\boldsymbol{\mu} \widetilde{\alpha} \widetilde{\boldsymbol{Z}} \boldsymbol{N}^{\widetilde{\alpha}-1}}_{\text{Labor demand}} \quad , \quad \widetilde{\boldsymbol{Y}} = \boldsymbol{\Omega} \widetilde{\boldsymbol{Z}} \boldsymbol{N}^{\widetilde{\alpha}} \quad , \quad \boldsymbol{C} = \left[1 - \frac{\delta}{R}(1-\gamma)\alpha\right] \frac{\widetilde{\boldsymbol{Y}}}{1-\alpha(1-\gamma)} .$$

With respect to an aggregate production function with productivity $\widetilde{\boldsymbol{Z}}$, the markdown $\boldsymbol{\mu}$ is a wedge that pushes the wage below the marginal product of labor, meanwhile for a given productivity $\widetilde{\boldsymbol{Z}}$ and employment $\boldsymbol{N}$, misallocation $\boldsymbol{\Omega}$ represents a direct reduction in output.[29] Note that the two terms appear *independently*.

Benchmark cases. Since welfare is determined by $\boldsymbol{C}$ and $\boldsymbol{N}$, and Proposition 1.2 allows us to restrict our attention to understanding markdowns $\boldsymbol{\mu}$ and misallocation $\boldsymbol{\Omega}$. Three benchmarks are useful:

- **Case I - Efficient allocation.** The efficient allocation coincides with an economy in which firm-by-firm wages and marginal revenue products of labor are aligned, that is $\mu_{ij} = 1$ for all firms. In this case $\boldsymbol{\mu} = 1$, and $\boldsymbol{\Omega} = 1$.

- **Case II - Monopsony limits.** A monopsonistically competitive economy attains under *either* of two limits: (1) $m_j \to \infty$ or (2) $\theta \to \eta$. Henceforth we simply refer to these conditions as the *"monopsony limits"*. Under either limit, firms are infinitesimal in the markets in which they set wages. In the first limit, they face a highly competitive local market. In the second limit, they face a national market. Markdowns μ_{ij} are identical across firms and equal to $\eta/(\eta+1)$, as market shares $s_{ij} \to 0$. In this case $\boldsymbol{\mu} = \mathbb{E}\left[\mu_{ij}\right]$, and $\boldsymbol{\Omega} = 1$.

- **Case III - Full model.** In our full model, the negative correlation of productivity and markdowns within markets (recall Figure 3), leads to (i) misallocation ($\boldsymbol{\Omega} < 1$), which reduces output, and (ii) a higher productivity weight on wide markdown firms, lowering $\boldsymbol{\mu} < \mathbb{E}\left[\mu_{ij}\right]$.

These special cases reveal that the oligopsonistic economy we have contributed distorts welfare relative to a monopsonistically competitive economy precisely through misallocation $\boldsymbol{\Omega}$. In a monoposonistically competitive economy, the labor supply elasticity to the firm η could be calibrated to generate the same $\boldsymbol{\mu}$, yet it would still feature $\boldsymbol{\Omega} = 1$. That $\boldsymbol{\Omega}$ is less than one is an outcome of the counterpart of both limits (i) labor markets are concentrated, and (ii) market power via $\theta < \eta$.

[29] Another way to see this is to define the following production function for competitive intermediate goods producers: $\widetilde{\boldsymbol{Y}} = \widetilde{\boldsymbol{Z}} \boldsymbol{N}^{\widetilde{\alpha}}$. The labor demanded by these producers is given by $\boldsymbol{W} = \boldsymbol{\mu} \widetilde{\alpha} \widetilde{\boldsymbol{Z}} \boldsymbol{N}^{\widetilde{\alpha}-1}$. A final goods producer with productivity $\boldsymbol{\Omega} < 1$ then converts intermediates into final goods.

This characterization of the model situates the remainder of our paper. First, we provide new empirical facts that allow us—along with the structure of the model—to credibly estimate θ and η. Second, we show that $\theta < \eta$ is necessary for the model to qualitatively and quantitatively match the sign and magnitude of non-targeted empirical micro-evidence on pass-through and strategic wage-setting of firms. Third, we show that the implied misallocation $\mathbf{\Omega}$ due to $\theta < \eta$ accounts for around half of the welfare losses due to labor market power, and that this is robust to specifications of aggregate preferences, even when aggregate labor supply is inelastic.

1.7 Measurement

The general equilibrium of the model can be used to show that the following two measures of the labor market are independent of the specification of the macroeconomy. We use these results in our calibration exercise in the next section.

Proposition 1.3. Labor share and concentration.

- *The* aggregate labor share *depends only on the distribution of markdowns and productivity*

$$LS_t := \frac{\int_0^1 \sum_{i=1}^{m_j} w_{ij} n_{ij}\, dj}{\int_0^1 \sum_{i=1}^{m_j} y_{ij}\, dj} = \frac{\boldsymbol{WN}}{\boldsymbol{Y}} = \gamma\alpha \left(\frac{\boldsymbol{\mu}}{\boldsymbol{\Omega}}\right)$$

- *The* across market payroll-weighted average of payroll concentration *is defined*

$$HHI_t^{wn} := \int_0^1 s_{jt}\, hhi_{jt}^{wn} dj \quad , \quad hhi_{jt}^{wn} = \sum_{i=1}^{m_j} s_{ijt}^2 \quad , \quad s_{jt} = \frac{\sum_{i\in j} w_{ijt} n_{ijt}}{\int_0^1 \sum_{i\in j} w_{ijt} n_{ijt}\, dj} \quad ,$$

- *The following holds, so HHI_t^{wn} depends only on the distribution of markdowns and productivity*

$$LS_t = \int_0^1 s_{jt} ls_{jt}\, dj = \underbrace{\alpha\gamma}_{\textit{Competitive LS}} \underbrace{\left[HHI_t^{wn} \left(\frac{\theta}{\theta+1}\right)^{-1} + \left(1 - HHI_t^{wn}\right)\left(\frac{\eta}{\eta+1}\right)^{-1}\right]^{-1}}_{\textit{Labor market power adjustment}} \tag{10}$$

For a full derivation of these results see Appendix E.5. Consider again the three benchmark cases. In an efficient economy, the labor share is equal to the output elasticity $\gamma\alpha$ and concentration plays no role. Under monopsony due to $m_j \to \infty$, the Herfindahl in each market is zero, all firms have the same markdown $\mu_{ij} = \boldsymbol{\mu} = \eta/(\eta+1)$, but with $\boldsymbol{\Omega} = 1$ the labor share is $\gamma\alpha\boldsymbol{\mu}$. Under monopsony due to $\theta \to \eta$, the Herfindahl in each market is positive but as firms compete nationally, drops out of the labor share. In our model, there is only an economically meaningful relationship between concentration and the labor share in concentrated, oligopsonistic markets with $\theta < \eta$.

In such an economy, higher concentration reduces the labor share. Intuitively, this expression arises in two steps. At the market level, the HHI measures the payroll share of high payroll share firms. In our model, these firms have wide markdowns and so low labor shares. Aggregating across firms within each market delivers (10) in the cross-section of markets. At the aggregate level, the aggregate labor share is the payroll weighted average of market labor shares, leading to (10).

Note that HHI_t^{wn} and LS_t are not sufficient statistics for welfare, even when combined with all other parameters of the model. Combined they reveal the ratio $(\boldsymbol{\mu}/\boldsymbol{\Omega})$, but cannot be used to disentangle the two. Proposition 1.2 established that both are required independently in order to compute aggregate quantities and hence welfare. Intuitively, the labor share and Herfindahl capture the wedge in the labor demand condition, but still do not capture the output wedge $\mathbf{\Omega}$.

Nonetheless, this model-implied measure of labor market concentration differs from all existing studies. For example, recent work by Benmelech, Bergman and Kim (2020) and Rinz (2018) use employment Herfindahls and various weighting schemes. Independent of our model framework, employment Herfindahls understate concentration since they ignore the positive relationship between wages and employment, which is a robust feature of the data (Charles Brown and James Medoff, 1989; Thierry Lallemand, Robert Plasman and François Rycx, 2007; Nicholas Bloom, Fatih Guvenen, Benjamin S. Smith, Jae Song and Till von Wachter, 2018).[30] We return to study the model's implication for the labor share through the lens of time-series of HHI_t^{wn} and equation (10) in Section 5.

2 Estimation

Our key parameters to estimate are the degree of across- (θ) and within- (η) market labor substitutability. In this section, we describe our novel approach which integrates (i) new empirical estimates from a quasi-natural experiment and (ii) new moments from the cross-section of markets, into (iii) a simulated method of moments routine in which all unknown parameters are estimated jointly. For additional moments see Table D1.

2.1 Approach - Structural vs. reduced-form labor supply elasticities

Structural elasticities. Our approach is motivated by the following observation. If a researcher could estimate the *structural elasticities of labor supply* that firms perceive at the Nash equilibrium level of employment, then they could combine data on payroll shares and one of the key model equations to estimate (θ, η):

$$\varepsilon\Big(s_{ij}, \theta, \eta\Big) := \left[\frac{\partial \log w_{ijt}}{\partial \log n_{ij}} \left(s_{ij}\right)\bigg|_{n^*_{-ij}}\right]^{-1} = \left[\frac{1}{\eta}\left(1 - s_{ij}^{wn}\right) + \frac{1}{\theta} s_{ij}\right]^{-1}. \tag{11}$$

In particular, a decreasing relationship between ε_{ij} and s_{ij} would identify $\eta > \theta$.

Reduced form elasticities. When firms behave strategically the structural elasticity cannot be measured using wage and employment responses to well identified firm-level shocks. As is clear from the notation above, the structural elasticity is a strictly partial equilibrium concept and answers the counterfactual: *How much will firm ij have to increase w_{ij} in order to expand n_{ij} by one percent, holding its competitors' employment fixed?* Given a shock to any firm in the market, an employment change at firm i will lead competitors to best-respond, which will cause i to best respond and so on. What an empiricist would measure in the data following a shock is therefore a *reduced-form elasticity* $\epsilon(s_{ijt}, \theta, \eta, \dots)$, which includes all other firms' employment and wage changes across market equilibria.[31]

[30] For a complete proof of this claim see Appendix E.

[31] We borrow the notation of ϵ for reduced-form elasticities and ε for structural elasticities from the estimation of structural macroeconomic models. In this literature *reduced-form shocks* which are empirical objects estimated out of VARs are often

Our insight is that, despite this, the reduced-form elasticities that we may aspire to measure, once filtered through our structural model, are still informative of (θ, η). To a first order approximation, the reduced-form elasticity of labor supply a researcher would measure for firm ij following a shock to it or a competitor is (for derivation see Appendix E.7):

$$\epsilon\Big(s_{ijt}, \theta, \eta, \dots\Big) := \frac{\mathrm{d}\log n_{ijt}}{\mathrm{d}\log w_{ijt}} = \left\langle \frac{1}{1+\varepsilon\Big(s_{ijt},\theta,\eta\Big)\left(\frac{\eta-\theta}{\theta\eta}\right)\left\{\sum_{k\neq i} s_{kjt}\frac{d\log n_{kjt}}{d\log n_{ijt}}\right\}} \right\rangle \times \varepsilon\Big(s_{ijt},\theta,\eta\Big). \tag{12}$$

A distinct property of (12) is that reduced form and structural elasticities coincide exactly under the monopsony limits. As $\theta \to \eta$, the term $\langle \cdot \rangle$ goes to one. As $s_{ijt} \to 0$, then the perturbed firm is infinitesimal so competitors do not respond and the equilibrium interaction term $\{\cdot\}$ goes to zero. Outside the monopsony limits, strategic interaction implies that reduced-form estimates of labor supply elasticities cannot be used to directly infer welfare-relevant labor supply elasticities, but are nonetheless indirectly informative as to parameter values when combined with a structural model.

Bias. The relationship between structural and reduced-form elasticities varies predictably depending on whether the underlying shock is idiosyncratic or common across multiple – but not all – firms in a market. A common shock to *all* firms drops out from the market equilibrium condition in Proposition 1.1 and could only be used to estimate the market level labor supply elasticity.

First, consider a positive idiosyncratic productivity shock to firm i in market j such that the firm expands employment. As the firm expands employment, its competitors respond. Since competition is Cournot, employment levels across firms are strategic substitutes so competitors reduce employment ($d\log n_{kjt} < 0$), implying that the equilibrium interaction term is negative, $\{\cdot\} < 0$, and the reduced-form elasticity exceeds the structural elasticity: $\epsilon(s_{ijt},\theta,\eta) > \varepsilon(s_{ijt},\theta,\eta)$. Figure 4A illustrates this case. The contraction in employment at competitors expands labor supply to the firm. An observer drawing conclusions about labor market power from the high reduced-form labor supply elasticity would conclude labor markets are more competitive than they are. In Section 2.6, we show that this bias is quantitatively significant: inferred structural and reduced-form elasticities differ by up to a factor of 2, even for perfectly idiosyncratic shocks.

For non-idiosyncratic shocks that are common across a subset of firms, we reach the opposite conclusion. Consider a tax cut that affects firm i in market j as well as the other large firms in market j. Call these affected firms *C-Corps*. Suppose the tax cut induces firm i and all affected C-Corps to expand employment, i.e. $d\log n_{ijt} > 0$ and $d\log n_{kjt} > 0$ for all firms $k \in j$ that are C-Corps. If non-C-Corp firms have small shares ($s_{kjt} \approx 0$), their strategic response is irrelevant. The equilibrium interaction term will be positive $\{\cdot\} > 0$, and the reduced-form elasticity understates the structural elasticity: $\epsilon\Big(s^{wn}_{ijt},\theta,\eta\Big) < \varepsilon\Big(s^{wn}_{ijt},\theta,\eta\Big)$. Figure 4B illustrates this case. The expansion in employment at competing C-Corps contracts labor supply to the firm. An observer drawing conclusions about labor market power from the low reduced-form labor supply elasticity would conclude that labor markets are less competitive than they are.

denoted ϵ, and *structural shocks* that are backed out of an estimated structural model are denoted ε.

Indirect inference. The above demonstrates that reduced-form elasticities are informative of structural elasticities which are in turn informative about welfare relevant parameters, and that the equilibrium structure of the model is necessary to complete this mapping. Our approach recognizes this. We first use a quasi-natural policy experiment to estimate the relationship between payroll shares and average *reduced-form* labor supply elasticities in the data: $\widehat{\epsilon}^{Data}(s)$. We then replicate the same policy experiment in our model which yields

$$\widehat{\epsilon}^{Model}\left(s,\theta,\eta\right) := \mathbb{E}\left[\epsilon^{Model}\left(s,\theta,\eta,\dots\right)\right],$$

where the expectation is being taken with respect to the distribution of all relevant labor market variables and shocks. We then choose (θ,η)—along with other parameters—such that when the quasi-natural experiment is simulated in the model, the model replicates our estimated empirical relationship between average reduced-form elasticities and payroll shares.

2.2 Estimating reduced-form labor supply elasticities in the data: $\widehat{\epsilon}^{Data}(s)$

We estimate size-dependent reduced-form labor supply elasticities using state corporate tax changes in conjunction with the Census Longitudinal Business Database (LBD).[32] The LBD provides high quality measures of employment, location, and industry with nearly universal coverage of the non-farm business sector. In order to proceed, we first define markets and firms. We then describe our regression approach.

Market. In our model, a *labor market* has two features: (i) a worker drawn at random from the economy will have a greater attachment to one labor market than others on the basis of idiosyncratic preferences, but will nonetheless be able to move across markets, and (ii) firms within a market compete strategically.

With these features in mind and given what we can observe in the LBD, we define a *local labor market* as a 3-digit NAICS (NAICS3) industry within a Commuting Zone (CZ).[33] Examples of adjacent 3-digit NAICS codes are subsectors 323-325: *'Printing and Related Support Activities'*, *'Petroleum and Coal Products Manufacturing'* and *'Chemical Manufacturing'* which we regard as suitably different. Examples of adjacent commuting zones include the collection of counties surrounding downtown Minneapolis and those surrounding Duluth.[34]

Firm. We define a firm in a local labor market as the collection of establishments operated by that firm. We aggregate employment and annual payroll of all establishments owned by the same firm within the same NAICS3-CZ market.[35] For each resulting firm-market-year observation, we observe employment, payroll, and herein define the *wage* as payroll per worker.

[32] We use the LBD (Bureau of the Census (2016*a*)) in conjunction with the SSEL (Bureau of the Census (2016*b*)) to identify C-Corporations as detailed in the appendix.

[33] Using BLS Occupational Employment Statistics microdata, Elizabeth Weber Handwerker and Matthew Dey (2018) show that when it comes to concentration there is little practical difference in defining a market at the occupation-city level rather than the industry-city level as these two measures are highly correlated. In particular, the across-city correlation of Herfindahl-Hirschman Indices at the CBSA-occupation and CBSA-industry level is 0.97.

[34] Many more examples are provided in Tables C2 and C3 in Appendix C.

[35] Firms are identified by the LBD variable *firmid*.

Regression framework. To estimate the reduced-form relationship $\widehat{\epsilon}^{Data}(s)$ in the data we use within-firm-market, across-time changes in wages and employment following state corporate tax changes.

Let i denote firm, j denote market, and t year. Let $s(j)$ denote the geographical state of market j. Let y_{ijt} denote a variable of interest at the firm-market-year level, such as employment or the wage. We are interested in coefficients on state corporate taxes $\tau_{s(j)t}$ and their interaction with lagged payroll shares. We use lagged payroll shares to avoid mechanical correlations between contemporaneous wages, employment and wage-bill shares, and control for lagged payroll shares s_{ijt-1} by themselves.[36] To isolate within-firm-market variation, we introduce firm-market fixed effects α_{ij}.[37] Lastly, as in Giroud and Rauh (2019) we include controls $X_{s(j)t}$ for the state unemployment rate and budget balance, along with a set of indicators for years in which state corporate income tax applied to gross receipts. Our regression specification is as follows:

$$\log y_{ijt} = \alpha_{ij} + \mu_t + \psi \, s_{ijt-1} + \beta^y \, \tau_{s(j)t} + \gamma^y \left(\tau_{s(j)t} \times s_{ijt-1} \right) + \Gamma X_{s(j)t} + \nu_{ijt}. \tag{13}$$

The coefficients β^y and γ^y capture the average effect of state corporate tax rate changes and their differential effect by market share. We estimate (13) separately for log employment and log wages (total payroll per worker). We then show how coefficient estimates from (13) can be used to construct $\widehat{\epsilon}^{Data}(s)$.

Clustering. We provide two sets of estimates which cluster at the state-year and market-year levels. Our estimated labor supply elasticity is a combination of both (i) the direct effect of taxes, and (ii) the interaction between payroll share and taxes. The former varies at the state-year level suggesting that clustering at the state-year level is appropriate; the latter varies at the firm-market-year level and since payroll shares contain market level variation, clustering at the market-year level is appropriate.

Sample. To abstract from changes in product market power we restrict our sample to tradeable industries identified by Mercedes Delgado, Richard Bryden and Samantha Zyontz (2014) and listed in Appendix C. Plants owned by the same firm are aggregated within a market, such that an observation is a firm-market-year. Since we rely on state-level corporate tax variation to generate changes in labor demand, we restrict our sample to C-Corporation firms (C-Corps) in the LBD from 1977 to 2011. Table C1 includes summary statistics of our 4.6 million observations at the firm-market-year level.

Estimates. Table 1 presents empirical estimates of (13). We start with (log) employment in year t as a dependent variable. Column (1) presents the full set of interaction terms between payroll shares and corporate taxes. Since $\tau_{s(j)t}$ is in units of percents, the coefficient on $\tau_{s(j)t}$ is an elasticity: a one percent increase in corporate taxes results in a 0.303 percent reduction in employment at firms that are atomistic within the market ($s_{ijt-1} = 0$). The interaction term is positive and significant. When combined with the negative direct effect, the interaction indicates a dampened response at larger firms. Compare the mean effect of a 1 ppt increase in $\tau_{s(j)t}$ on a firm with a mean payroll share (0.03) to a firm with a one standard deviation higher share (0.10).[38] Employment declines by -0.27 percent at the small firm and

[36] State-level corporate taxes are proportional flat-taxes on firms' accounting profits. Our data for state-level corporate taxes comes from the data made publicly available by Giroud and Rauh (2019): https://web.stanford.edu/~rauh/.

[37] In this exercise only, we exclude commuting zones that straddle multiple states.

[38] Among tradeable markets with at least two firms, the wage-bill weighted mean payroll share is 0.030 with a standard deviation of 0.063

		Year t		Year $t+1$	
		$\log n_{ijt}$	$\log w_{ijt}$	$\log n_{ijt+1}$	$\log w_{ijt+1}$
		(1)	(2)	(3)	(4)
State corporate tax	$\tau_{s(j)t}$	-0.00303***	-0.00244***	-0.00258***	-0.00120**
		(0.000672)	(0.000702)	(0.000768)	(0.000604)
		[0.000331]	[0.000287]	[0.000355]	[0.000297]
Payroll share	s_{ijt-1}	0.967***	0.0805***	0.763***	0.0727***
		(0.0304)	(0.00960)	(0.0261)	(0.0102)
		[0.0110]	[0.00617]	[0.0114]	[0.00664]
Interaction	$\tau_{s(k),t} \times s_{i,j,k,t-1}$	0.0119***	0.00492***	0.0118***	0.00390***
		(0.00319)	(0.00128)	(0.00282)	(0.00138)
		[0.00134]	[0.000791]	[0.00143]	[0.000845]
Fixed Effects		Y	Y	Y	Y
R-squared		0.907	0.780	0.888	0.730
Firm-market-year observations		4.26m	4.26m	4.26m	4.26m

Table 1: Estimation results for equation (13)

Notes: All specifications include fixed effects for: (i) year, (ii) firmid×market. According to Census requirements, the number of observations is rounded to the nearest 1,000. Standard errors in round parentheses $(\cdot)$ are clustered at State × Year level. Standard errors in square parentheses $[\cdot]$ are clustered at Market × Year level. Sample includes tradeable C-Corps from 1977 to 2011.

	A. Year t elasticities				
	$\widehat{\epsilon}^{Data}(s_{ijt-1})$	p-value: $\epsilon^{Data}(s_{ijt-1}) = 0$		p-value: $\epsilon^{Data}(s_{ijt-1}) = \epsilon^{Data}(0)$	
		state-year	market-year	state-year	market-year
	(1)	(2)	(3)	(4)	(5)
1% payroll share, $s_{ijt-1} = 0.01$	1.2200	0.0002	0.0000	0.1024	0.0006
5% payroll share, $s_{ijt-1} = 0.05$	1.1120	0.0009	0.0000	0.0998	0.0005
10% payroll share, $s_{ijt-1} = 0.10$	0.9462	0.0085	0.0000	0.0978	0.0003
	B. Year $t+1$ elasticities				
	$\widehat{\epsilon}^{Data}(s_{ijt-1})$	p-value: $\epsilon^{Data}(s_{ijt-1}) = 0$		p-value: $\epsilon^{Data}(s_{ijt-1}) = \epsilon^{Data}(0)$	
		state-year	market-year	state-year	market-year
	(1)	(2)	(3)	(4)	(5)
1% payroll share, $s_{ijt-1} = 0.01$	2.1210	0.0221	0.0003	0.4472	0.1613
5% payroll share, $s_{ijt-1} = 0.05$	1.9780	0.0487	0.0014	0.4308	0.1439
10% payroll share, $s_{ijt-1} = 0.10$	1.7240	0.1350	0.0131	0.4043	0.1154

Table 2: Elasticities and hypothesis testing

Notes: Panel A Column (1) constructs elasticities based on the Date t estimates in Columns (1) and (2) in Table 1 using equation (14). Column (2) reports p-value of the hypothesis test $H0 : \epsilon^{Data}(s) = 0$ using standard error clustered at the state-year level. Column (3) clusters at the market-year level. Column (4) reports p-value of the hypothesis test $H0 : \epsilon^{Data}(s) = \epsilon^{Data}(0)$ using standard error clustered at the state-year level. Column (5) clusters at the market-year level. Panel B repeats the exercise based on the Date $t+1$ estimates in Columns (3) and (4) in Table 1.

-0.18 percent at the large firm. Consistent with Giroud and Rauh (2019), increases in corporate tax rates reduce employment. Our new empirical finding is that this reduction is nearly 40 percent weaker at larger firms.

Column (2) illustrates estimates of (13) when the dependent variable is the wage. Qualitatively the signs echo the employment response: on average wages fall, and this decline is smaller at larger firms. Columns (3) and (4) provide estimates of (13) using year $t+1$ employment and wages as dependent variables. These specifications are designed to accommodate adjustment frictions in prices and quantities. We again find a negative effect of corporate taxes on employment and wages, with diminished effects at larger firms.

Share-dependent reduced-form labor supply elasticities. Table 2 combines the wage and employment responses to compute the relationship between the average reduced-form labor supply elasticity and payroll shares, which inform θ and η. Differentiating (13) with respect to $\tau_{s(j)t}$ delivers share-dependent reduced-form wage and employment elasticities:

$$\frac{\mathrm{d}\log n_{ijt}}{\mathrm{d}\tau_{s(j)t}} = \beta^n + \gamma^n s_{ijt-1}\ ,\ \frac{\mathrm{d}\log w_{ijt}}{\mathrm{d}\tau_{s(j)t}} = \beta^w + \gamma^w s_{ijt-1}\ ,\ \widehat{\epsilon}^{Data}(s_{ijt-1}) = \frac{\widehat{\mathrm{d}\log n_{ijt}}}{\widehat{\mathrm{d}\log w_{ijt}}} = \frac{\widehat{\beta}^n + \widehat{\gamma}^n s_{ijt-1}}{\widehat{\beta}^w + \widehat{\gamma}^w s_{ijt-1}} \tag{14}$$

When we turn to indirect inference, we run the same regressions on model simulated data to compute $\epsilon^{Model}(s)$ in the same way.

Column (1) of Table 2A reports reduced-form labor supply elasticity estimates $\widehat{\epsilon}^{Data}(s_{ijt-1})$ based on the Table 1 Year t estimates for $s_{ijt-1} \in \{0.01, 0.05, 0.10\}$. At a wage bill share of 1 percent, the year t reduced-form labor supply elasticity is 1.22, and declines to 0.95 at a wage bill share of 10 percent. Columns (2) and (3) show that the elasticity is statistically significant at the 5 percent level under either assumption for clustering.

In Columns (4) and (5) of Table 2A, we test whether the estimated date t labor supply elasticities of larger firms are statistically different from atomistic firms. Formally, we test $H_0 : \widehat{\epsilon}^{Data}(s_{ijt-1}) = \widehat{\epsilon}^{Data}(0)$ for $s_{ijt-1} \in \{0.01, 0.05, 0.10\}$. For wage-bill shares of 5% and 10%, the year t reduced-form labor supply elasticities in column (1) are significantly different from that of an atomistic firm at the 10 percent level. Table 2B repeats the same exercise for year $t+1$ employment and wage responses from columns (3) and (4) of Table 1. At year $t+1$ the implied reduced-form labor supply elasticities are larger, potentially due to slow employment adjustment. However, the estimates are noisier.

In summary, our more precise year t estimates of the size-dependent wage and employment response indicate (i) less responsiveness of larger firms, and (ii) significantly lower reduced-form labor supply elasticities of larger firms. Our year $t+1$ estimates imply greater labor supply elasticities across all firm sizes, consistent with frictional adjustment. In both cases we find that larger firms have lower labor supply elasticities; however, we lack the power to statistically distinguish the labor supply elasticity of large firms from small firms in the year $t+1$ case.

Additional results. Appendix G.3 provides additional results. First, estimation of our structural model simply requires *consistent* auxiliary moments that can be simulated. The threat to *consistency* when we estimate equation (13) is that forces we will not models move employment and wages at the state-year level (e.g. taxes are cut when unemployed is low). Table G3 shows that our main interaction between corporate taxes and the wage-bill share is robust to the inclusion of state-year fixed effects, thus removing all common state-year variation. Second, we directly compute the ratio of wage changes to employment changes at the firm-level and study their relationship with firms' wage-bill share. Following corporate tax cuts, we estimate statistically significantly different labor supply elasticities at large relative to atomistic firms. Third, using the 2012 Census of Manufacturers, we show that variation in non-wage compensation is unable to explain the large movements in markdowns implied by our baseline labor supply elasticity estimates. Finally, we show that systematic variation in capital intensity by market share cannot explain our results: within markets, capital intensity and payroll shares are only weakly correlated.

2.3 Simulating reduced-form labor supply elasticities in the model: $\widehat{\epsilon}^{Model}(s,\theta,\eta)$

To construct $\widehat{\epsilon}^{Model}(s,\theta,\eta)$, we add corporate taxes to the environment and show how they shift marginal revenue products of labor. We make several modifications to our theory. Corporate taxes are a tax on profits, net of interest payments on debt. Firms finance $\lambda_K \in [0,1]$ of their capital using debt and maximize post-tax profits:

$$\pi_{ijt} = \Big(1-\tau_C\Big) z_{ijt} \Big(k_{ijt}^{1-\gamma} n_{ijt}^{\gamma}\Big)^{\alpha} - \Big(1-\tau_C\lambda_K\Big) R k_{ijt} - \Big(1-\tau_C\Big) w(n_{ijt}) n_{ijt},$$

A random fraction $\omega_C \in [0,1]$ of firms in each market are C-corps and subject to τ_C; all other firms face $\tau_C = 0$. In the data, C-corps are larger on average. To capture this we assume a productivity premium $\Delta_C > 1$:

$$\log(z_{ijt}) \sim \begin{cases} N\Big(1,\sigma_z^2\Big) & \text{if } i \text{ is not a } C\text{-corp (i.e. with probability } 1-\omega_C) \\ N\Big(\Delta_C,\sigma_z^2\Big) & \text{if } i \text{ is a } C\text{-corp (i.e. with probability } \omega_C) \end{cases}$$

For C-corps, the corporate tax distorts their capital decision, which reduces the marginal product of labor. Under the firm's optimal capital demand, effective productivity $\widetilde{z}_{ijt}$ is decreasing in τ_C if $\lambda_K > 0$:

$$\frac{\pi_{ijt}}{1-\tau_C} = \max_{n_{ijt}} \widetilde{z}_{ijt} n_{ijt}^{\widetilde{\alpha}} - w(n_{ijt}) n_{ijt} \quad , \quad \widetilde{z}_{ijt} = \left[\frac{1-\tau_C}{1-\tau_C\lambda_K}\right]^{\frac{\alpha(1-\gamma)}{1-\alpha(1-\gamma)}} \times \left\langle \left[1-\alpha\left(1-\gamma\right)\right] \left(\frac{\alpha\left(1-\gamma\right)}{R_t}\right)^{\frac{\alpha(1-\gamma)}{1-\alpha(1-\gamma)}} z_{ijt}^{\frac{1}{1-\alpha(1-\gamma)}} \right\rangle,$$

see Appendix E.9. With these modifications to the theory, we can simulate an increase in corporate taxes and estimate reduced-form elasticities consistent with our approach to the data to obtain $\widehat{\epsilon}^{Model}(s,\theta,\eta)$.

2.4 Indirect inference

We estimate the model using 2014 Census data and proceed in two steps. First, to match the reduced-form elasticities measured in a sample of tradeable firms, we estimate a tradeable-only version of our economy. This includes the corporate tax experiment and yields estimates of the key preference parameters η and θ. Second, holding η and θ fixed, we drop the corporate tax experiment and estimate the remaining parameters to match economy-wide moments. The tradeable sector is more concentrated than the economy on the whole, so the second step is necessary for measuring labor market power in the US economy. We add to the model a parameter that shifts firm productivity $\overline{Z}$, and a preference parameter that shifts labor supply $\overline{\varphi}$.

Common external parameters. On an annual basis, the discount rate is 4 percent ($\beta = 0.9615$), and the depreciation rate is 10 percent ($\delta = 0.10$). Throughout we simulate $5,000$ markets and verify that our results are not sensitive to this choice. The moments used in our estimation are robust to alternative specifications of aggregate preferences $U(\boldsymbol{C}, \boldsymbol{N})$, so we defer specifying U until we evaluate welfare.

Tradeable only - External parameters. To capture the distribution of tradeable firms across markets, $m_j \sim G(m_j)$ we combine a Pareto distribution with a discrete mass at $m_j = 1$ to capture single firm markets. The mass of tradeable markets with a single firm is 16 percent (Table F1). We fit the remaining Pareto parameters to match the first three moments of the distribution of firms across markets. Appendix Table F1 provides moments and parameter estimates.

The fraction of capital financed by debt is chosen to match the debt to capital ratio among tradeable firms. For this we use tradeable firms in Compustat and obtain $\lambda_K = 0.213$. We compute that 31 percent of firm-market observations are C-corps, and so set $\omega_C = 0.31$.

Given parameters we simulate a three-period panel from the model. The first two periods are given by the model's steady state with τ_C set to the mean state corporate tax rate of 6.9 percent (the average over the sample period 1977 to 2011). In the third period, we increase taxes by Δ_τ equal to one percentage point. This is approximately one standard deviation of the distribution of state corporate tax changes observed in our sample period.[39] Treating model output as panel data, we estimate exactly our empirical specification (13) with firm fixed effects and lagged payroll shares (hence the requirement for three periods). We replicate our treatment of the data, and transform point estimates into average reduced-form elasticities by payroll share using equation (14). Appendix F includes additional details on simulating the tax experiment.

Tradeable only - Estimated parameters. We now estimate $\boldsymbol{\psi} = \{\theta, \eta, \gamma, \alpha, \sigma_z, \Delta_C, \overline{Z}, \overline{\varphi}\}$. A key element of our strategy is to use moments that are independent of aggregate preferences. This allows us to conduct robustness checks with respect to aggregate preferences without recalibrating the model.

To estimate θ and η, we target the reduced-form labor supply elasticities in Table 2. Year t and $t+1$ elasticities have different merits. Year t elasticities are less likely to include confounding factors, whereas date $t+1$ elasticities alleviate concerns regarding adjustment frictions. As a compromise we target the average of year t and year $t+1$ estimates. Rather than targeting the entire function (14), we compute the average reduced-form labor supply elasticity of firms with payroll shares between 0 and 5 percent, and 5 and 10 percent. This captures the bulk of variation in our data. The value for small firms is most informative of η. The value for large firms is most informative of θ. We estimate $\theta = 0.42$ and $\eta = 10.85$.

We estimate productivity dispersion σ_z to match the payroll weighted wage-bill Herfindahl of 0.17 (Table D1). Increasing σ_z increases the market power of large firms, increasing concentration. We pin down α and γ using the capital and labor share of income.[40] As can be seen from equation (10), conditional on HHI_t^{wn}, α shifts the labor share. Figure A2 shows how in practice this argument pins down α

[39] We use data made publicly available by Giroud and Rauh (2019)

[40] We use BEA data to compute the tradeable labor share of 53.9 percents. The remaining non-labor income is apportioned according to the share of capital and profits in the aggregate economy. The aggregate capital share is 18 percent based on Simcha Barkai (2020). Apportioning yields a tradeable capital share of 19.3 percent.

Parameter		Value	Moment	Model	Data
I. **Tradeable industries only**					
$G(m_j)$	Pareto and point mass at $m_j = 1$		Mean, Variance, Skewness of distribution 15 percent of markets have 1 firm		
ω_C	Share of firms that are C-corps	0.31	Share of estabs. that are C-corps (CBP, 2014)		
τ_C	State corporate tax rate	0.069	Mean of state corp. tax rate $\tau_{C,st}$		
Δ_τ	State corporate tax rate increase	0.010	Std. dev. of annual $\tau_{C,st}$		
λ_K	Fraction of capital debt financed	0.213	Tradeable industries (Compustat, 2014)		
Estimated					
θ	Across market substitutability	0.42	Average $\hat{\varepsilon}^{Data}(s)$ for $s \in [0.05, 0.10]$	1.49	1.43
η	Within market substitutability	10.85	Average $\hat{\varepsilon}^{Data}(s)$ for $s \in [0, 0.05]$	1.53	1.61
Δ_C	Relative productivity of C-corps	1.40	Employment share of C-corps	0.63	0.64
σ_z	Productivity dispersion	0.186	Payroll weighted $\mathbb{E}[hhi_j^{wn}]$	0.17	0.17
α	Decreasing returns to scale	0.992	Labor share	0.53	0.54
γ	Exponent on labor	0.797	Capital share	0.19	0.19
$\overline{Z}$	Productivity shifter	1.72×10^4	Mean firm size	34.6	34.6
$\overline{\varphi}$	Labor disutility shifter	2.171	Mean worker earnings (\$000)	58.3	58.3
II. **All industries**					
$G(m_j)$	Pareto and point mass at $m_j = 1$		Mean, Variance, Skewness of distribution 9 percent of markets have 1 firm		
θ	Across market substitutability	0.42	Held fixed at estimated tradeable value		
η	Within market substitutability	10.85	Held fixed at estimated tradeable value		
Estimated					
σ_z	Productivity dispersion	0.312	Payroll weighted $\mathbb{E}[hhi_j^{wn}]$	0.11	0.11
α	Decreasing returns to scale	0.940	Labor share	0.57	0.57
γ	Exponent on labor	0.808	Capital share	0.18	0.18
$\overline{Z}$	Productivity shifter	1.79×10^4	Mean firm size	22.8	22.8
$\overline{\varphi}$	Labor disutility shifter	3.099	Mean worker earnings (\$000)	43.8	43.8

Table 3: Summary of Parameters

and σ around the estimated parameters. Our estimate of α implies close to constant returns to scale in the tradeable sector. Parameter γ matches the aggregate capital share. Table 3 summarizes all parameters and the model's fit to the target moments. In Appendix E.6 we provide a closed form solution of the model and prove that in any equilibrium $(\overline{Z}, \overline{\varphi})$ normalize units of wages and labor, so are chosen to match average firm size (34.6) and payroll per worker (\$58,300) (Table D1). Finally, Δ_C matches the 64 percent employment share of C-corps which we compute in Census data.[41]

Economy-wide calibration. Holding our estimates of preference parameters η and θ fixed, we recalibrate our model to match economy-wide moments. We update the distribution of firms across markets $G(m_j)$, which almost halves the number of markets with one firm to 9 percent. We remove C-corps, setting $\omega_C = 0$ and estimate $\boldsymbol{\psi} = \{\gamma, \alpha, \sigma_z, \overline{Z}, \overline{\varphi}\}$ to match, the (i) labor share, (ii) capital share, (iii) payroll-weighted wage-bill Herfindahl, (iv) average firm size, and (v) average payroll per worker. Notably, in the overall economy, concentration is lower, the labor share is higher, wages are lower, and average firm size is smaller. With less concentration market power is lower, reducing profits, hence a lower value of α

[41]We construct this statistic directly from our data, which compares closely with the statistic available from the SUSB of 65.9%.

is required to increase profits. The model matches the data exactly and yields decreasing returns to scale, $\alpha = 0.94$, and dispersion of log productivity, $\sigma_z = 0.312$, which is consistent with values that appear in the firm dynamics literature.

2.5 Discussion of estimated θ and η

Figure 4C plots $\widehat{\epsilon}^{Data}(s_{ijt-1})$ over $s_{ijt-1} \in [0, 0.10]$. The model generates a downward sloping reduced-form labor supply elasticity similar to the data. Notably, the reduced-form estimates for atomistic firms are roughly five times smaller than the structural estimates. Thus, a naive inference, based on reduced form elasticities alone, would conclude that the labor market is less competitive than it actually is, and infer wide markdowns at atomistic firms of 0.65. Our structural estimates of the labor supply elasticity at atomistic firms imply a markdown of only 0.92, roughly 3 times narrower. When filtered through the model, the data implies a more competitive labor market than one would assess from taking the reduced form elasticity estimates at face value. The upward bias in market power implied from naive use of reduced-form estimates is less pronounced among larger firms. These predictions are in line with our theory of non-idiosyncratic shocks in Panel B.

Entry and Exit. One concern may be that following a tax increase some firms may exit, and this may affect our estimates of θ and η. To address this we conduct two exercises. First, in Appendix G we estimate a linear probability model of firm-market exit in year $t+1$ as a function of corporate taxes in year t. We find economically insignificant results. Giroud and Rauh (2019) find that firms adjust their *total number of plants* in the state. Our results imply that they do not appear to be exiting commuting-zone markets *entirely* in response to corporate tax changes. Second, despite these insignificant empirical results, we estimate the model under the extreme counterfactual assumption that the smallest 5 percent of C-corps in each market exit after the tax increase. Our estimates of θ and η are unchanged. Details of this exercise are in Appendix G.2.

2.6 Implied bias when using idiosyncratic shocks to measure labor supply elasticities

Many existing papers rely on estimation strategies that assume atomistic firms and infer monopsony from firm-level responses to idiosyncratic shocks (e.g., see the articles surveyed in Card et al., 2018). So far, our analysis has focused on bias between reduced-form and structural elasticities when the observed shock is non-idiosyncratic, illustrated in Figure 4B. Does our theory pose issues for inferring labor supply elasticities when the identifying variation is purely idiosyncratic? In this section, we quantify such bias under idiosyncratic shocks, illustrated in Figure 4A, and discussed in Section 2.1. We show that using data on employment and wage changes in response to identified firm-level shocks to infer key structural parameters may generate sizeable bias due to the violation of SUTVA.

To quantify this bias, we ask what one would infer from data generated by a truly idiosyncratic shock to a single randomly selected firm in our economy. We draw a firm at random, increase its productivity, recompute the market Nash equilibrium, and calculate the reduced form elasticity $\widehat{\epsilon}_{ij}$ off of the firm's

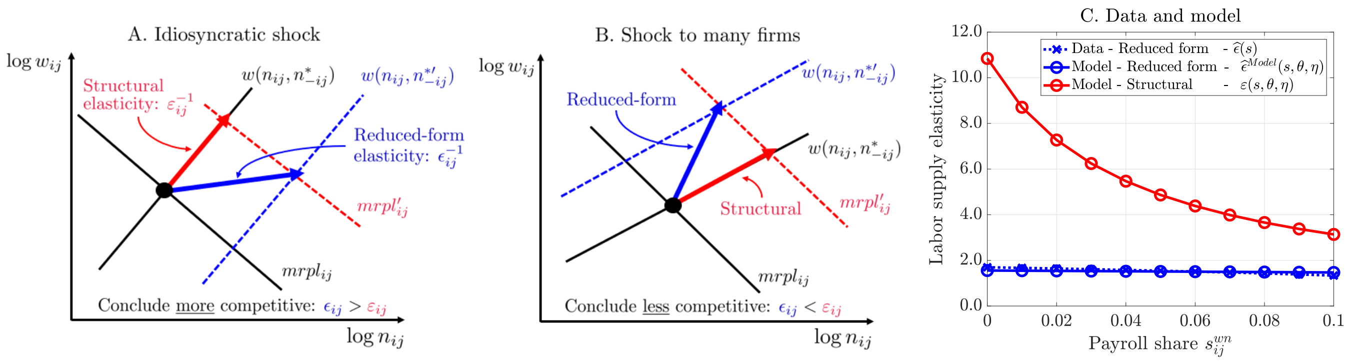

Figure 4: Reduced form and structural labor supply elasticities

Notes: **Panel A** demonstrates the relationship between reduced-form and structural labor supply elasticities following an idiosyncratic shock. **Panel B** demonstrates the relationship between reduced-form and structural labor supply elasticities for a non-idiosyncratic shock that affects a proper subset of firms. **Panel C** compares reduced-form and structural labor supply elasticities by firm payroll share in response to a corporate tax shock of 1 percentage point 'Data - Reduced form' is an equally weighted average of the date $t+1$ and date t empirical labor supply elasticity estimates in Table 2. 'Model - Reduced form t' plots reduced-form labor supply elasticity estimates, estimated on simulated model data as described in Appendix F.2. 'Model - Structural' plots $\varepsilon(\cdot)$ from equation (11).

employment and wage changes. We compare this to the structural elasticity ε_{ij}. We repeat this 5,000 times for small (one percent), and large (50 percent) productivity shocks and plot the results in Figure 5. This Monte Carlo exercise reveals a significant difference in reduced-form and structural labor supply elasticities for firms with market shares not equal to 0 or 1, even when the identifying variation is a *perfectly idiosyncratic* shock to firm labor demand. The bias between reduced-form and structural elasticities goes up to 200 percent for firms with market shares between 0 and 10 percent. Accounting for the oligopolistic equilibrium of the local labor market is quantitatively important for recovering welfare relevant parameters, even when the supposed identifying variation is entirely idiosyncratic.

This exercise implies that even if a researcher aims to use ideal idiosyncratic variation in labor demand to infer structural elasticities and do welfare analysis, they would have to deflate their *reduced-form elasticity* estimates in order to recover the true *structural elasticities*. Absent this adjustment one would conclude labor supply elasticities are larger, which would lead one to infer narrower markdowns and conclude that the macroeconomic effects of labor market power are lower. The details of our Monte Carlo exercise are included in Appendix F.3.

Figure 5 shows that two important caveats apply, both summarized in equation (12). If the firm has a share of one, then reduced-form and structural elasticity estimates coincide and reveal θ. If the firm has an infinitesimal share, then reduced-form and structural elasticity estimates coincide and reveal η. Finally, a *market level* shock will directly reveal θ, so long as the market itself is not large. If the market is very large then a market level shock will also effect the macroeconomic equilibrium of the labor market, and reduced-form elasticities will be contaminated by φ.

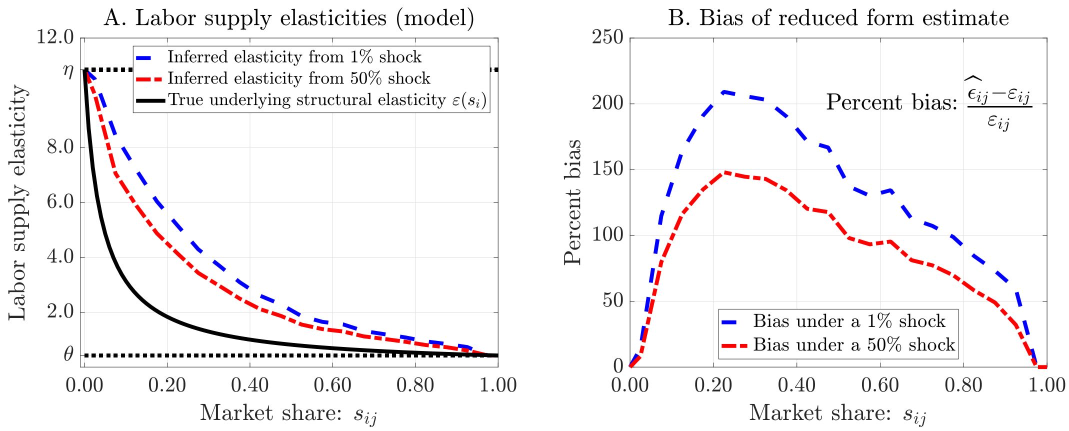

Figure 5: Reduced form and structural elasticities in response to idiosyncratic productivity shocks.

Notes: Panel A plots Monte Carlo results which compare reduced-form to structural labor supply elasticities in response to a perfectly idiosyncratic shock to a single firm. The lines labeled 'Reduced form elasticity' plot the average estimated reduced-form labor supply elasticity $\hat{\varepsilon}(s)$ as detailed in Appendix F.3. The dashed line labeled 'Structural elasticity' plots $\varepsilon(s)$ from equation (11). Panel B reports the error of the average reduced-form elasticity relative to the structural elasticity: $100 \times (\hat{\varepsilon}(s) - \varepsilon(s))/\varepsilon(s)$.

3 Validation

In this section we show that our oligopsony model with $\theta < \eta$ is qualitatively and quantitatively consistent with independent evidence by comparing the model's implications for (i) pass-through of value added to wages, and (ii) strategic responses of firms to competitors' wage changes. In each case we show how the monopsony limits of our model ($\theta = \eta$, or $M_j = \infty$) qualitatively and quantitatively fail, while our estimated model matches the data. A summary is as follows:

1. **Pass-through** - Under $\theta = \eta$, pass-through from value added per worker to wages is equal to one. Kline et al. (2019) produce an estimate of 0.47. In a replication of their exercise the model produces an estimate of 0.50
2. **Competitor responses** - Under $\theta = \eta$, a firm's response to the wage increase of a competitor will be close to zero. Staiger, Spetz and Phibbs (2010) produce an estimate of 0.128. In a replication of their exercise the model produces an estimate of 0.109.

While (1.) could be generated by an alternative model with non-homothetic preferences or non-isoelastic production technology and monopsonistic competition, (2.) is a direct test of the oligopoly model in this paper.[42] Figure A1 shows that the model replicates the distribution of markets by concentration, both unweighted and payroll weighted. Consistent with the data, Table A2 shows that (i) concentration measured in terms of employment is lower than concentration measured in terms of payroll, and (ii) unweighted measures of concentration are 2 to 3 times larger than when weighted across markets.[43]

[42] In related work, David W. Berger, Kyle F. Herkenhoff and Simon Mongey (2021*b*) shows that the model also replicates the cross-employer wage elasticities estimated by Ellora Derenoncourt, Clemens Noelke and David Weil (2021) following wage increases at Amazon. Since this experiment is closely related to Staiger, Spetz and Phibbs (2010), we do not include it here.

[43] In the data (model) weighted average concentration measured in terms of employment is 0.15 (0.16) and in terms of payroll is 0.17 (0.17). In the data (model) unweighted average concentration measured in terms of employment is 0.45 (0.32) and in terms of payroll is 0.48 (0.33).

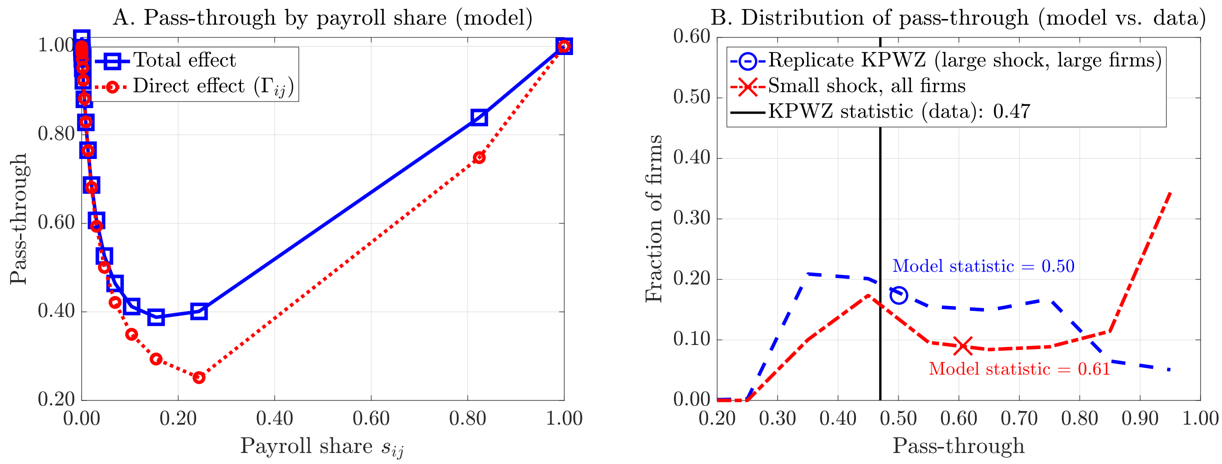

Figure 6: Pass-through and replication of Kline et al, (2018)

Notes: Panel A computes average pass-through in bins by 20 ventiles of the payroll share distribution. We draw one firm from each market at random and increase its productivity by 1 percent. We resolve the market equilibrium, keeping general equilibrium aggregates fixed. Within each bin we compute the mean of $\Delta \log w_i / \Delta \log vapw_i$ of these firms (blue solid line, squares) We use equation (16) and compute Γ_{ij} for each firm based on initial market shares, and again take averages within each bin (red dotted line, circles). The histogram in Panel B plots the fraction of firms with firm-level pass-through $\Delta \log w_i / \Delta \log vapw_i$ in bins of width 0.10, both with (blue, circles) and without (red, crosses) the size restrictions imposed to match the sample statistics of KPWZ. The "Small shock, all firms" case, considers a productivity shock of 1 percent, the KPWZ replication has a shock of 19 percent.

3.1 Pass-through - Kline et al. (2019)

Theory. A body of recent empirical evidence documents that the elasticity of worker wages with respect to value added per worker following shocks to firm productivity is less than one (Kline et al., 2019; Card et al., 2018). Under our theory, equation (7) implies:

$$\underbrace{w_{ijt} = \alpha\gamma \times \mu_{ijt} \times vapw_{ijt}}_{\text{A. Levels}} \quad , \quad \underbrace{\Delta \log w_{ijt} = \Delta \log \mu_{ijt} + \Delta \log vapw_{ijt}.}_{\text{B. Log changes}} \tag{15}$$

The literature discusses pass-through in two ways, in levels as in (15)A. or in log changes as in (15)B. Imperfect pass-through in the first case could be due to markdowns, $\mu < 1$, or decreasing returns to scale, $\alpha\gamma < 1$. Imperfect pass-through in the second case, however, is a prediction of our model. In order for log wages to respond less than one-for-one with changes in log value added per worker, e.g. $\Delta \log w_{ijt} < \Delta \log vapw_{ijt}$, markdowns must increase. Our oligopsony model naturally generates this variation in markdowns: following an increase in firm productivity firms hire more workers, pay higher wages, but with an expanding market share the firm's markdown widens, which dampens their wage increase. In either monopsony limit, markdowns are constant and $\Delta \log w_{ijt} = \Delta \log vapw_{ijt}$, as in monopsony models of Manning (2003) and Card et al. (2018).

Totally differentiating the market equilibrium system yields the following first order approximation for pass-through following any perturbation (for derivation see Appendix E.8):

$$\underbrace{\frac{\Delta \log w_{ij}}{\Delta \log vapw_{ij}}}_{\text{Pass-through}} = \underbrace{\Gamma^*_{ij}}_{\text{Direct}} + \underbrace{(1-\Gamma^*_{ij}) \sum_{k \neq i} \frac{s^*_{kj}}{1-s^*_{ij}} \frac{\Delta \log w_{kjt}}{\Delta \log vapw_{ijt}}}_{\text{Indirect}} \quad , \quad \Gamma^*_{ij} = \frac{s^*_{ij}(\eta-\theta)+\theta(\eta+1)}{[1+(1+\eta)(1-s^*_{ij})]s^*_{ij}(\eta-\theta)+\theta(\eta+1)}, \tag{16}$$

where Δ's are taken with respect to the initial equilibrium, which is denoted by asterisks. Clearly from this expression under either monopsony limit ($\theta \to \eta$ or $s_{ij} \to 0$) $\Gamma^*_{ij} = 1$ and so pass-through is one.

Figure 6 plots the average of the direct effect Γ^*_{ij} and total effect which combines the response of the firm to the responses of competitors, following a 1 percent productivity shock to an individual firm.[44] As firms become larger in a market, two offsetting forces shape pass-through. First, the direct effect declines, as increases in productivity go into increasing market power and widening markdowns, reducing pass-through. Second, when the firm is large its competitors respond more aggressively, increase their wages, which indirectly leads to further wage increases at the firm, increasing pass-through. Hence the total effect implies more pass-through than the direct effect. On net, the direct effect dominates and pass-through is less than one.

Replication. Estimates for wage pass-through from a paper with sufficient details for us to replicate come from Kline et al. (2019).[45] KPWZ exploit patent issuance as an instrument, comparing consequent changes in value added per worker and wages. To replicate their quasi-experiment we solve the model in general equilibrium, then draw one firm from each market and increase their productivity by ψ_1^{KPWZ} percent. We solve the new market equilibria, keeping aggregates constant. We limit our sample of firms that we shock to firms with more than ψ_2^{KPWZ} workers. We calibrate the replication parameters $\{\psi_1^{KPWZ}, \psi_2^{KPWZ}\}$ to match two moments of their study: a mean firm size in sample of 61.83, which is larger than in our baseline calibration, and a mean increase in value added per worker relative to mean value added per worker of 13 percent.[46] Table A1 compares summary statistics of our regression sample to theirs.

Measurement. To measure pass-through, we adopt the procedure in KPWZ.[47] We treat the pre- and post- observations from the model as a panel with two observations per firm. We then regress w_{it} on $vapw_{it}$ in levels with a firm-specific fixed effect. The regression coefficient is a semi-elasticity which is

[44] The direct effect can be computed in closed form, but the total effect requires simulating the model since it depends on the distribution of competitor shares within the market.

[45] Recent work by Card et al. (2018) uses lagged log sales per worker as an instrument for log value added per worker. From Table 2 (panel A, row IV, column 1) their estimate of pass-through is 32.7 percent, however the paper contains insufficient information in order for us to replicate it, for example the size of changes in value added per worker. A structural approach is taken by Benjamin Friedrich, Lisa Laun, Costas Meghir and Luigi Pistaferri (2019), who estimate pass-through of 31 percent from permanent shocks in a model of worker and firm dynamics estimated on Swedish employer-employee data (Table 12, column 1).

[46] See KPWZ. We take the *Mean firm size* of 61.83 from their Table II, panel A, column 5. The percentage increase in VAPW is 0.13=15.74/120.16, where 15.74 is the mean increase in value added per worker (Table V, column 4), and 120.16 is the mean value added per worker (Table II, panel A, column 5). *Value-added* in KPWZ is defined as sales minus 'costs of goods sold net of labor costs'. This is consistent with our measure.

[47] They describe this procedure in Section VII, and footnotes to Table VIII.

converted into an elasticity using the initial period mean wage and initial period mean value added per worker (see their Section 7). With this procedure their point estimate implies pass-through of 0.47, with a standard error of 0.23. We verify that under monopsony, i.e. $\theta = \eta$, this approach delivers a pass-through of one in the model.

Results. Figure 6B provides the results of this exercise. Replicating the KPWZ statistic, our estimate of pass-through is 0.50, less than on seventh of a standard deviation above the their estimate. We view this as a success of the model. Using our model we can go beyond this point estimate and plot the distribution of pass-through across firms, showing rich cross-sectional heterogeneity. Doing so presents two important considerations for future empirics. First, relatively smaller firms have higher pass-through, and the support of firm pass-through extends below the KPWZ estimate. Measuring pass-through at any one firm would give different results depending on the firm's market share. Second, if we were to ignore that the KPWZ sample is (i) biased toward large firms and (ii) studies a large shock, the pass-through statistic would increase by around 10 percentage points (red cross) as smaller firms have higher pass-through, and pass-through is smaller for large shocks.[48] Measuring pass-through is sensitive to the sample of firms and shock used as a source of variation.

3.2 Strategic responses - Staiger, Spetz and Phibbs (2010)

Theory. An important paper by Staiger, Spetz and Phibbs (2010, henceforth SSP) provides direct empirical evidence regarding the response of firms in one labor market to increases in wages of other firms in the same labor market. Consider either monopsony limit where a firm exogenously narrows its markdown to $\mu' \in [\mu, 1]$, where $\mu = \eta/(\eta+1)$. In either limit, the fact that the firm is infinitesimal implies that this would have zero effect on competitor's wages within the same geographic area. Contrary to this, SSP find that when Department of Veterans Affairs (VA) hospitals increased their wages due to a change in policy, competitors increased their wages in response. In an environment with $\eta > \theta$, the above pass-through formula (16) shows how our model is consistent with this fact as an increase in wage at firm $k \neq i$ causes firm i to increase its wage. The mechanism is as follows: a VA hospital increases its wage, which increases its employment and increases its market share $s_{VA,jt}$, this tightens competition leading non-VA hospitals to narrow their markdowns, which increases their wages.

Replication. Key properties of the sample and quasi-experiment in SSP are as follows: (i) markets—defined as a 15-mile radius of the focal VA hospital—had on average 10.9 hospitals, (ii) the VA hospital was on average paying nurses 1.9 percent below the average wage for nurses at non-VA hospitals, (iii) the policy increased nurse wages of VA hospitals paying below the local average up to the average wage of nurses at non-VA hospitals. To replicate this experiment we take our baseline economy which we call period zero. We then isolate markets j with between 9 and 13 firms, draw one firm i at random in each of these markets from the set of firms with a wage w_{ij0} between 1 and 3 percent less than the average market wage, and then increase this firm's productivity by ψ_{ij}^{SSP} percent. Holding aggregates fixed, we

[48] Decreasing returns is not behind these results. Conducting the exercise under $\alpha = 1$ increases pass-through only slightly.

	Model	Data
A. Replication statistics		
Average log difference (gap) between VA hospital wage and average competitor wage	0.020	0.019
Average number of firms in a market	10.8	10.9
Average productivity increase to set gap to zero		1.1
B. Result		
Elasticity of competitor wages to VA hospital wage	0.109	0.128
(Standard error)		(0.033)

Table 4: Strategic interaction and replication of Staiger, Spetz and Phibbs (2010)

Notes: Model simulation selects firms (the 'VA hospital') whose wages are between 1% and 3% lower than the average market wage and are in a market with 9 to 13 firms. The exercise is to raise the VA hospital wage in period one up to the average market wage in period zero, and then to compute the response to competitor wages. Pooling across markets, we report a cross-sectional elasticity obtained by regressing log changes of average competitor wages on log changes of VA hospital wages. We compare our estimates to Table 1 (summary statistics) and Table 2 (point estimates) in Staiger, Spetz and Phibbs (2010).

then solve the new market equilibria. We choose ψ_{ij}^{SSP} firm-by-firm such that in the new equilibrium the wage w_{ij1} at firm i equals the initial period average wage at competitors.[49] On average ψ_{ij}^{SSP} is 5.21 percent, and ranges from 2.30 percent to 8.53 percent.

Measurement. To measure employer wage responses, we adopt the procedure in SSP. We treat the data from the model as a panel with two periods. From this we compute $\Delta \log w_{VA,j}$ at the 'VA hospital' in each market, and the change in log wages at non-*VA* hospitals $\Delta \log w_{ij}$. We then pool across markets and estimate regression equation (6) of SSP which produces a coefficient α_1 comparable to their Table 2, column 1:

$$\Delta \log w_{ij} = \alpha_0 + \alpha_1 \Delta \log w_{VA,j} + e_{ij}$$

Results. Table 4 compares our results to SSP. Quantitatively, the model generates a response of competitors' wages of 10.9 percent, which is within one standard deviation of the SSP estimate of 0.128. We conclude that the structure of labor markets and our estimates of θ and η generate strategic complementarities in concentrated labor markets that are consistent with this important empirical evidence.

Summary. In summary, our model is shown to provide a quantitative foundation for key recent empirical studies that document incomplete pass-through of changes in productivity to changes in wages, and the responses of competitors to firms changes in their wage.

4 Implications of labor market power

Now that we have validated our model quantitatively, we use it to measure labor market power in the U.S. at the micro- and macro-level and explore it's implications for macroeconomic outcomes: output and welfare.

[49] An alternative would have been to have narrowed the VA hospital's markdown. From the perspective of the competing firms, both are equivalent, since they only take into account competitor's wages.

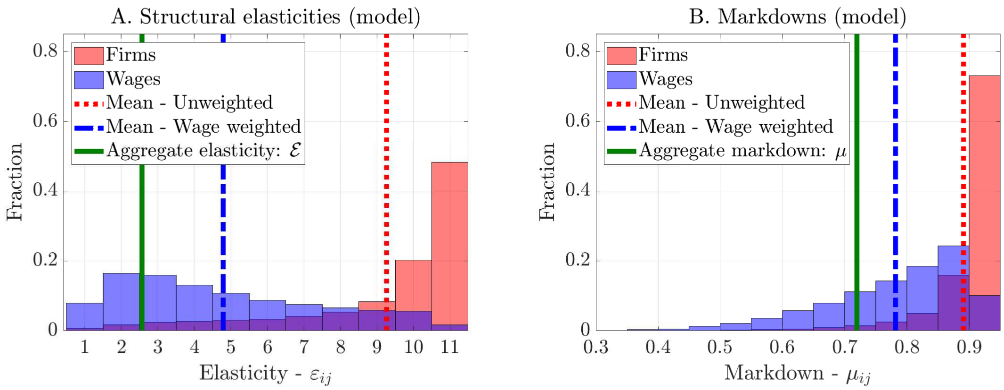

Figure 7: Distribution of labor market power across firms

Notes: Panel A plots the distribution of equilibrium structural labor supply elasticities $\varepsilon(\cdot)$ from equation (11), unweighted ('Firms') and weighted by payroll ('Wages'), $\mathcal{E}$ is the aggregate structural labor supply elasticity consistent with an aggregate markdown $\boldsymbol{\mu}$, i.e. $\boldsymbol{\mu} = \mathcal{E}/(\mathcal{E}+1)$. Panel B conducts the same exercise for markdowns, where $\boldsymbol{\mu}$ is the aggregate markdown.

4.1 Microeconomic measurement

A firm's markdown is an economically meaningful measure of labor market power. The markdown at the firm measures the wedge between the firm's marginal revenue product of labor and its wage. In an efficient economy, workers are paid their entire marginal revenue product and thus markdowns are equated to one.

Figures 7A and 7B plot the distribution of firms and wage payments across structural labor supply elasticities ε_{ij} and markdowns μ_{ij}. In an economy that matches the distribution of firms and concentration across markets, as well as salient pass-through and wage setting facts, we find that most firms in the economy are highly competitive, with narrow markdowns attributable to low market shares and high labor supply elasticities. Taking an unweighted average across firms, the mean labor supply elasticity is more than nine, while the markdown is more than 10 percent $(\mathbb{E}[\mu_{ij}] = 0.89)$.

Despite this, the distribution of wage payments in the economy is highly skewed toward firms with more labor market power. Weighted by payroll, the average labor supply elasticity drops to less than five, and the average markdown is around 0.78. As we have shown in Section 1.6, however, what matters for welfare is the aggregate markdown $\boldsymbol{\mu}$. This is a particular productivity weighted average of firm markdowns, and skews even further, with a value of 0.72. To provide context for this number we compute what we call the *representative labor supply elasticity*, which is the elasticity of labor supply to a firm that would lead a firm to set a markdown of $\boldsymbol{\mu}$. This value $\mathcal{E}$ is around 2.57, which is less than a third of the cross-sectional average of ε_{ij}.

In short, the distribution of wage payments in the economy is crucial for determining the mapping from labor supply elasticities to the wedges that determine output, employment and aggregate wages. Labor market power functions *as if* there were a single firm facing a labor supply elasticity that is less than three, despite most payroll being at firms with labor supply elasticities closer to 5 and most firms having a labor supply elasticity closer to 10.

4.2 Macroeconomic measurement

In the Appendix we define an efficient allocation. The efficient allocation can be decentralized under a *competitive equilibrium* concept in which firms take their wages as given, such that $\mu_{ijt} = 1$. With all markdowns equal to one the aggregate markdown $\boldsymbol{\mu}^* = 1$ and misallocation $\boldsymbol{\Omega}^* = 1$.

We measure the *welfare loss / gain* across steady-states, which we denote λ, as the percentage increase in consumption in the benchmark economy, that would be required to make the household indifferent with respect to a counterfactual allocation.[50] Let $\{\boldsymbol{C}, \boldsymbol{N}\}$ denote consumption and disutility of labor in the benchmark economy and $\{\boldsymbol{C}^*, \boldsymbol{N}^*\}$ consumption and disutility in the efficient economy. Then λ equates $U\big((1+\lambda)\,\boldsymbol{C}, \boldsymbol{N}\big) = U\big(\boldsymbol{C}^*, \boldsymbol{N}^*\big)$. For our baseline results we consider GHH preferences, and then introduce wealth effects (WE):

$$U_{GHH}(\mathbf{C}_t, \boldsymbol{N}_t) = \log\left(\boldsymbol{C}_t - \overline{\varphi}^{-1/\varphi}\frac{\boldsymbol{N}_t^{1+\frac{1}{\varphi}}}{1+\frac{1}{\varphi}}\right) \quad , \quad U_{WE}(\mathbf{C}_t, \boldsymbol{N}_t) = \frac{\boldsymbol{C}_t^{1-\sigma}}{1-\sigma} - \overline{\varphi}^{-1/\varphi}\frac{\boldsymbol{N}_t^{1+\frac{1}{\varphi}}}{1+\frac{1}{\varphi}}.$$

We show in Appendix E that $\boldsymbol{\mu}$ and $\boldsymbol{\Omega}$ are also independent of the scale parameters $\overline{Z}$ and $\overline{\varphi}$. Across comparative statics we recalibrate $(\overline{Z}, \overline{\varphi})$ to match the same average worker wage and average firm payroll in the benchmark oligopsony economy.

Results. Table 5 presents our baseline results. We focus on an aggregate Frisch elasticity of 0.50, which we vary within the range considered by the Congressional Budget Office in assessing policy: $\varphi \in [0.20, 0.80]$. First, steady state welfare gains are 7-8 percent of consumption.[51] Second, under a higher labor supply elasticity, it less costly to supply more labor in the competitive allocation, and so welfare and output gains are larger. Third, leveraging Proposition 1.2 we decompose these effects into the aggregate markdown $\boldsymbol{\mu}$ and misallocation $\boldsymbol{\Omega}$. Roughly half is due to $\boldsymbol{\Omega}$, a third due to $\boldsymbol{\mu}$, and the remainder due to their interaction.[52] Fourth, in terms of measurable aggregates, the average worker wage increases by about 40 percent, with employment changes ranging from 2 to 22 percent, again increasing in φ. Fifth, concentration more than doubles. Absent labor market power, wages and employment increase most at the largest firms. These firms had the widest markdowns in our benchmark economy. This reduces misallocation (i.e. $\boldsymbol{\Omega}$ rises) but also increases concentration.[53]

Reallocation. A key result is the significant role played by misallocation, $\boldsymbol{\Omega}$. Recall that a monopsony economy would deliver $\boldsymbol{\mu} = \eta/(\eta+1)$ and $\boldsymbol{\Omega} = 1$. Hence the single preference parameter of a monopsony economy, η, could be calibrated to match $\boldsymbol{\mu}$. Such an economy would have the same macroeconomic

[50] Note that aggregate consumption incorporates the effect of competition on wages, employment and firm profits. Recall that $\boldsymbol{W}$ is defined by $\boldsymbol{WN} = \int \sum_{i\in j} w_{ij} n_{ij}\, dj$, and $\boldsymbol{C}$ is defined by $\boldsymbol{C} = \int \sum_{i\in j} c_{ij}\, dj$. Therefore, aggregating firms' profit conditions ($\pi_{ij} = y_{ij} - w_{ij}n_{ij} - Rk_{ij}$) under goods market clearing and these definitions returns the household budget constraint ($\Pi = \boldsymbol{C} - \boldsymbol{WN} - RK$), so $\boldsymbol{C} = \Pi + \boldsymbol{WN} + RK$.

[51] In a previous version of this paper we showed that these gains are moderated by around 1 percent when taking into account transition dynamics (Berger, Herkenhoff and Mongey, 2021*a*). Reaching higher steady-state capital is costly and gradual due to decreasing marginal utility. Transition dynamics are straight-forward to compute, since Proposition 1.2 tells us that $\boldsymbol{\mu}_t$ and $\boldsymbol{\Omega}_t$ jump at date zero to their efficient levels.

[52] As an example, for the first of these calculations we keep $\boldsymbol{\mu}$ fixed at the level associated with the oligopsony economy, and then set the other wedge $\boldsymbol{\omega}$ equal to the efficient benchmark $\boldsymbol{\omega}^* = 1$. We then use our set of six general equilibrium conditions (Proposition 1.2) to recompute the aggregate economy under $(\boldsymbol{\mu}, \boldsymbol{\omega}^*)$, and the associated increase in welfare.

[53] Consistent with Proposition 1.3, this increase in concentration is independent of the specification of aggregate preferences.

Frisch elasticity	A. Consumption equivalent welfare gains			B. Aggregates			C. Concentration
	Steady state	% Due to $\boldsymbol{\Omega}$	% Due to $\boldsymbol{\mu}$	Output	Ave. wage	Employment	Weighted
φ	$\lambda \times 100$	λ^{Ω}/λ	λ^{μ}/λ	% change	% change	% change	ΔHHI^{wn}
0.20	5.7	77.6	17.6	10.9	44.5	2.1	0.23
0.50	7.6	57.8	33.1	20.9	43.5	12.2	0.23
0.80	9.6	45.8	42.4	31.4	42.6	22.7	0.23

Table 5: Benchmark welfare gains from competition - GHH preferences

Notes: Changes are measured from benchmark oligopsony economy to competitive equilibrium.

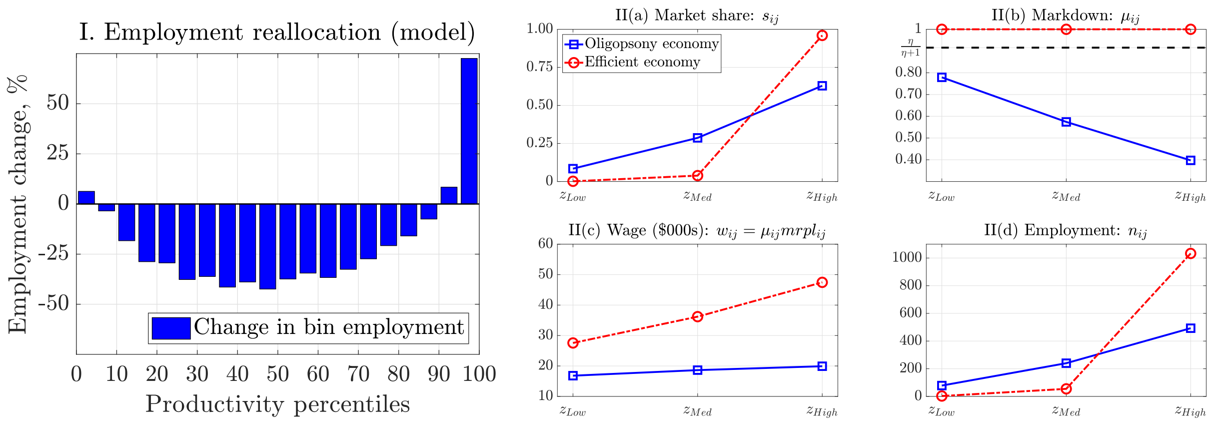

Figure 8: Employment reallocation from the benchmark economy to the efficient economy

Notes: As in Figure 3, low, medium and high productivities of the firms correspond to the 10^{th}, 50^{th} and 90^{th} percentiles of the productivity distribution.

implications of markdowns shifting to their efficient level. However, it would miss the large macroeconomic effects of resolving misallocation.

Figure 8(I) illustrates the reallocation of employment that underlies the increase in $\boldsymbol{\Omega}$ as labor market power is dissipated. Panel I shows the significant shift in employment away from low productivity firms and toward the highest decile of firms. To visualize reallocation at the market level, Panel II returns to the three firm example in Figure 3 and adds the efficient allocation. With all markdowns equal to one, wages increase at all firms, more than doubling at the most productive firm. Since it had the widest markdown to begin with, the wage increase is largest at the most productive firm. This reallocates employment away from the medium and low productivity firms, increasing market $\boldsymbol{\omega}_j$.

Wealth effects. We can learn more about the macroeconomic effects of misallocation and markdowns by studying the economy with wealth effects. Less labor market power reduces misallocation and increases consumption, which leads to a reduction in household labor supply, dampening output effects. Despite this, our results are robust. Compared to our baseline output effects of 20 percent, Figure 9A shows that shifting to $U_{WE}(\cdot)$ under log preferences ($\sigma = 1$), output losses due to labor market power are lower. Further increases in σ as far as four reduce output losses, but losses are still significant, around 5 percent.

Interestingly, wealth effects have a significant impact on the decomposition of welfare gains into

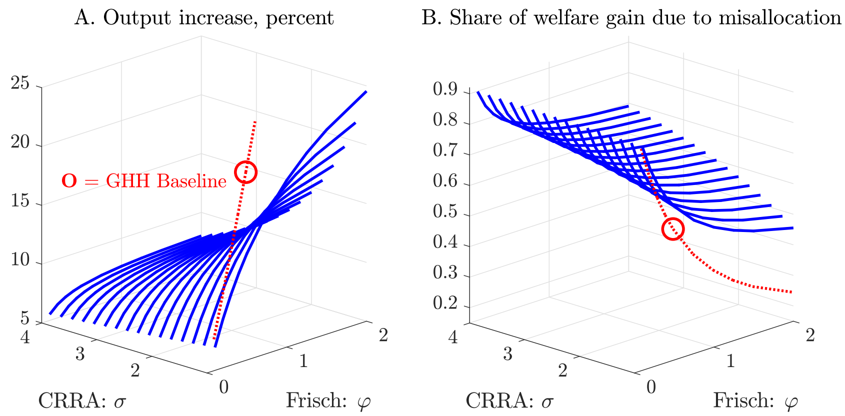

Figure 9: Welfare costs of labor market power with wealth effects

Notes: Each economy, indexed by (σ, φ), has the same concentration, average firm size, average worker wage, and all other moments used in our calibration Table 3. Panel B plots the percent of Panel A welfare gains due to increasing $\boldsymbol{\omega}$ to $\boldsymbol{\omega}^* = 1$.

misallocation and markdowns. Figure 9B shows that the fraction of welfare gains attributable to the resolution of misallocation $\boldsymbol{\Omega}$ jumps to nearly 60 percent under log preferences, and increases as wealth effects become more pronounced. Recall that in general equilibrium $\boldsymbol{Y} = \frac{1}{1-(1-\gamma)\alpha}\boldsymbol{\Omega}\widetilde{\boldsymbol{Z}}\boldsymbol{N}^{\widetilde{\alpha}}$, and $\boldsymbol{C} \propto \boldsymbol{Y}$. Hence an increase in $\boldsymbol{\Omega}$ delivers a direct increase in consumption. If σ is larger, the increase in utility due to the increase in consumption is larger, and so the share of welfare gains due to resolving misallocation are greater.

A key result is that even with fixed labor supply, so no role for $\boldsymbol{\mu}$, labor market power can have large output effects. In the limit as $\varphi \to 0$ labor supply becomes perfectly inelastic and $\boldsymbol{\Omega}$ accounts for the entirety of the welfare gains from competition.[54] That is, even when labor supply is completely inelastic in the aggregate, there are still macroeconomic effects of labor market power due to its microeconomic implications for the allocation of labor across productive units in all markets of the economy.

Entry. In these exercises we have kept the set of firms in the economy fixed. We leave to future research the complicated question of how entry interacts with a comparative static with respect to market structure. David Rezza Baqaee and Emmanuel Farhi (2020*a*) provides a starting point. That paper studies the role of entry in understanding the output effects of firm-level productivity shocks. They study inefficient economies with imperfectly substitutable goods, decreasing returns in production and fixed markups. Our environment has imperfectly substitutable jobs, decreasing returns in production and endogenous markdowns. The key, complicated steps to extend this to our setting include (i) endogenizing markdowns, (ii) taking a stand on the directedness of entry across markets.

5 Application- Labor market concentration and labor's share, 1977 - 2013

As an application of our framework, we use the model implied relationship between concentration and the labor share to show how alternative measurements of concentration can lead to different counterfac-

[54] In the case of GHH preferences and $\varphi \to 0$, the household is hand-to-mouth, and employment is fixed regardless of $\boldsymbol{\mu}$.

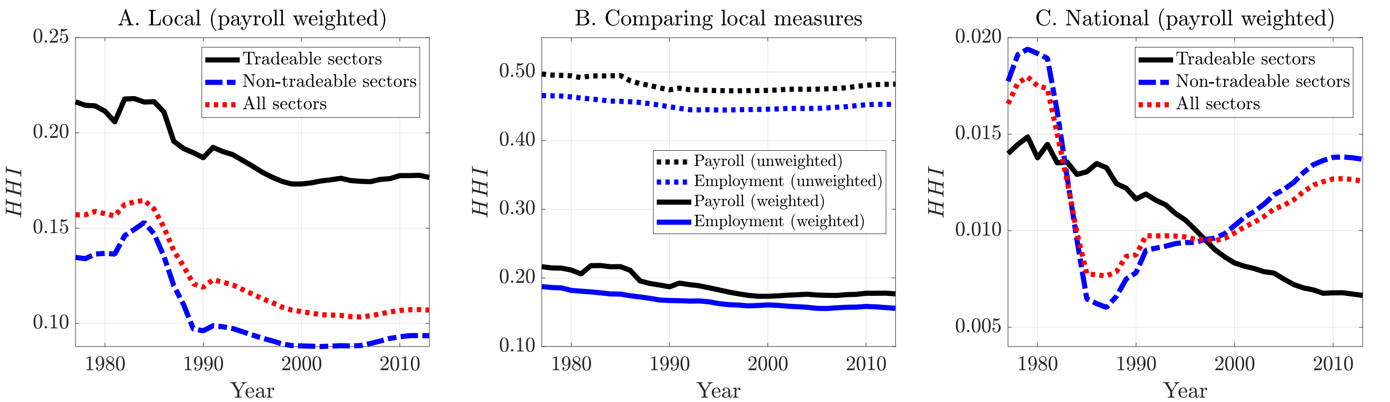

Figure 10: Measures of labor market concentration, 1977 to 2013

Notes: Data is plotted using a centered 5-year moving average in all panels. Panel A plots the payroll weighted average of the wage-bill Herfindahl computed at the commuting zone × NAICS3, $HHI_t^{wn} = \int_0^1 s_{jt} hhi_{jt}^{wn}\, dj$. Panel A includes three lines for tradeables (NAICS2 codes of 11,21,31,32,33,55), non-tradeables (all other NAICS2 codes), and the whole economy. Panel B compares the tradeable payroll weighted and unweighted CZ×NAICS3 wage-bill Herfindahl: hhi_{jt}^{wn}. Panel B also compares the employment weighted and unweighted CZ×NAICS3 employment Herfindahl: hhi_{jt}^{n}. Panel C plots the national payroll weighted wage-bill Herfindahl. National Herfindahls are computed at the NAICS3 level, ignoring geography, then weighted by industry payroll.

tual predictions. We leverage the model's mapping from concentration to the labor share:

$$LS_t = \int_0^1 s_{jt} ls_{jt}\, dj = \alpha\gamma \left[HHI_t^{wn} \left(\frac{\theta}{\theta+1} \right)^{-1} + \left(1 - HHI_t^{wn} \right) \left(\frac{\eta}{\eta+1} \right)^{-1} \right]^{-1}, \; HHI_t^{wn} = \int_0^1 s_{jt} hhi_{jt}^{wn}\, dj. \quad (17)$$

The model clearly implies a welfare relevant measure of labor market concentration: payroll weighted wage-bill Herfindahl. Figure 10A shows how this has evolved from 1977 to 2013, using our definition of a *local* labor market: a 3-digit industry and commuting zone.[55] Tradeables, non-tradeables, and the combined economy all decline over time. In tradeables the decline is roughly 20 percent from 0.217 to 0.175. Concentration in non-tradeables is lower, and declines with a slight increase at the end of our sample, but by 2013 is half its level in 1984.

Figure 10B demonstrates the importance of weighting and compares payroll and employment concentration, considering only the tradeable sector.[56] First, not weighting across markets inflates the measure of concentration by a factor of around 2.5 for both payroll and employment concentration. Many markets have few employers but they account for a very small fraction of wage payments. Second, the weighted payroll and employment Herfindahls display similar trends, with a time-series correlation of 0.75 between 1977 and 2013. Despite this, the positive size wage premium leads employment concentration to be 20 percent less than payroll concentration.

Figure 10C repeats this exercise disregarding the local nature of labor markets. We first compute concentration at the national industry level, and then weight across industries. According to this measure, which is irrelevant for welfare, labor market concentration increased over this period, following a

[55] To meet Census disclosure requirements, we show detailed summary statistics in 1976 and 2014 in Appendix D. Our time series graphs cover the complementary years from 1977 to 2013.

[56] We have been unable to disclose the corresponding statistics for non-tradeable sectors.

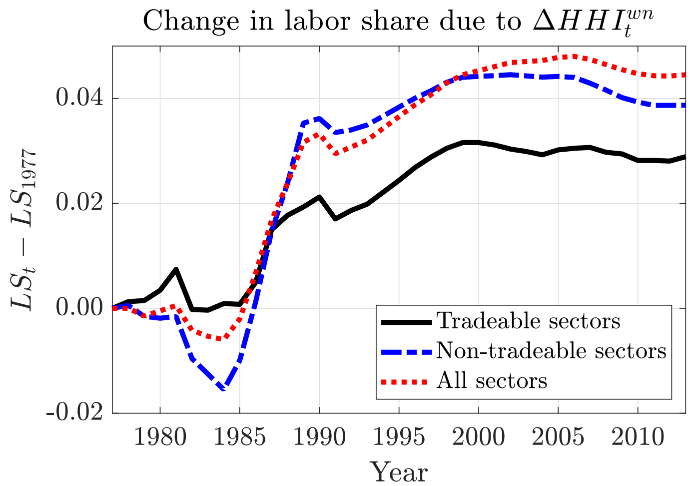

Figure 11: Change in labor share attributable to change in payroll Herfindahl, 1977 to 2013

Notes: Figure constructed by using estimates of payroll weighted wage-bill Herfindahl (Figure 10A) and the expression for labor's share of income (17). $\{\gamma, \alpha, \eta, \theta\}$ held fixed at values in Table 3.

sharp drop in the early 1980s. While our payroll Herfindahl measure is distinct, other contemporaneous work has documented a disconnect between national and local employment Herfindahls using different definitions of markets and aggregation (e.g. Rossi-Hansberg, Sarte and Trachter, 2018; Rinz, 2018).[57]

5.1 Counterfactual labor share, 1977 - 2013

We can now combine three of the novel contributions of this paper to link the dynamics of labor's share of income to labor market power: (i) the closed-form expression for labor's share of income given by equation (17), (ii) our estimates of θ and η, and (iii) our new time-series of aggregate concentration (Figure 10).

Our counterfactual holds $\{\gamma, \alpha, \eta, \theta\}$ fixed and varies the payroll weighted wage-bill Herfindahl HHI_t^{wn} from 1977 to 2013, using this to compute the implied labor share from equation (17). At our estimated parameters, the declining wage-bill Herfindahl between 1977 and 2013 contributed to increase the labor share by around 4 percentage points. Figure 11 plots the implied changes in labor share holding all else fixed except for the payroll weighted wage-bill Herfindahl. The predicted upward pressure of declining local Herfindahls on labor's share of income is similar for tradeables, non-tradeables, and the overall economy. We conclude that changes in labor market concentration are unlikely to have contributed to the declining labor share in the United States (e.g. Loukas Karabarbounis and Brent Neiman, 2013).

[57] First, Rinz (2018) describes employment concentration in a number of non-tradeable sectors using a NAICS4×Commuting zone definition of a labor market. Second, Rinz (2018) does not aggregate establishments within firms when computing employment shares at the local level. When averaged at the 2-digit level, he finds similar trends in tradeable and non-tradeable sectors.

6 Conclusion

We measure oligopsony in administrative U.S. Census data through the lens of a structural model. By doing so, we make several contributions. We develop a general equilibrium model of labor market oligopsony that combines differentiation of jobs via preference heterogeneity and concentrated labor markets. We prove that the model is block recursive and provide a closed-firm link between labor market concentration and labor's share of income. We show how to estimate the underlying preference parameters that govern labor market power in the presence of strategic interactions. We provide novel measures of firm size-dependent labor supply elasticities. We rationalize empirical evidence suggestive of oligopsony by quantitatively replicating two empirical papers. A monopsony version of our model does not replicate these studies. Under a variety of aggregate preferences, we compute output losses of 5 to 20 percent from labor market power. These provide upper bounds on the welfare effects of policies that might mitigate labor market power. We leave to future work how these may be affected by additional considerations such as skill heterogeneity and entry and exit. We show that roughly half of the gains are attributable to misallocation by using a novel representative agent counterpart of our economy. Lastly, we show that the model relevant measure of concentration is the payroll weighted wage-bill Herfindahl, which we measure, and use to show that changes in labor market concentration are unlikely to have contributed to a falling labor share in the U.S.

We believe our framework and empirical findings provide many avenues for future research. By establishing the empirical relevance of our framework through validation tests, we provide the literature with a useful point of departure. In ongoing work, we demonstrate the framework can be modified to replicate empirical studies in the merger and minimum wage literatures, and then be used to contribute to debates on merger and minimum wage policy (Berger, Herkenhoff and Mongey, 2021*b*). The model can also incorporate firm entry/exit and worker heterogeneity, accommodating use of occupation or worker-level data such as the Longitudinal Employer Household Dynamics database to estimate oligopsony.

References

Amiti, Mary, Oleg Itskhoki, and Jozef Konings. 2014. "Importers, Exporters, and Exchange Rate Disconnect." *American Economic Review*, 104(7): 1942–78.

Amiti, Mary, Oleg Itskhoki, and Jozef Konings. 2019. "International Shocks, Variable Markups, and Domestic Prices." *The Review of Economic Studies*, 86(6): 2356–2402.

Anderson, Simon P., Andre De Palma, and Jacques-Francois Thisse. 1987. "The CES is a Discrete Choice Model?" *Economics Letters*, 24(2): 139–140.

Ashenfelter, Orley C, Henry Farber, and Michael R Ransom. 2010. "Labor Market Monopsony." *Journal of Labor Economics*, 28(2): 203–210.

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

Azar, José, Ioana Marinescu, and Marshall I. Steinbaum. 2020. "Labor Market Concentration." *Journal of Human Resources*, 1218–9914R1.

Azar, José, Ioana Marinescu, Marshall Steinbaum, and Bledi Taska. 2020. "Concentration in US labor markets: Evidence from online vacancy data." *Labour Economics*, 66: 101886.

Azkarate-Askasua, Miren, and Miguel Zerecero. 2020. "The Aggregate Effects of Labor Market Concentration." Working Paper.

Banfi, Stefano, and Benjamin Villena-Roldan. 2018. "Do high-wage jobs attract more applicants? Directed search evidence from the online labor market." *Journal of Labor Economics*, 37(3).

Baqaee, David Rezza, and Emmanuel Farhi. 2020*a*. "Entry vs. Rents: Aggregation with Economies of Scale." National Bureau of Economic Research NBER Working Paper 27140.

Baqaee, David Rezza, and Emmanuel Farhi. 2020*b*. "Productivity and Misallocation in General Equilibrium." *The Quarterly Journal of Economics*, 135(1): 105–163.

Barkai, Simcha. 2020. "Declining Labor and Capital Shares." *The Journal of Finance*, 75(5): 2421–2463.

Becker, Gary S. 1962. "Investment in human capital: A theoretical analysis." *Journal of Political Economy*, 70(5, Part 2): 9–49.

Belot, Michele, Philipp Kircher, and Paul Muller. 2017. "How Wage Announcements Affect Job Search Behaviour-A Field Experimental Investigation." European University Institute Working Papers.

Benmelech, Efraim, Nittai Bergman, and Hyunseob Kim. 2020. "Strong Employers and Weak Employees: How Does Employer Concentration Affect Wages?" *Journal of Human Resources*, 0119–10007R1.

Berger, David W., Kyle F. Herkenhoff, and Simon Mongey. 2021*a*. "Labor Market Power." Federal Reserve Bank of Minneapolis Opportunity and Inclusive Growth Institute Working Papers 48.

Berger, David W., Kyle F. Herkenhoff, and Simon Mongey. 2021*b*. "Minimum Wages and Welfare." *Working Paper*.

Bhaskar, Venkataraman, Alan Manning, and Ted To. 2002. "Oligopsony and monopsonistic competition in labor markets." *Journal of Economic Perspectives*, 16(2): 155–174.

Bloom, Nicholas, Fatih Guvenen, Benjamin S. Smith, Jae Song, and Till von Wachter. 2018. "The Disappearing Large-Firm Wage Premium." *AEA Papers and Proceedings*, 108: 317–22.

Boal, William M, and Michael R Ransom. 1997. "Monopsony in the Labor Market." *Journal of Economic Literature*, 35(1): 86–112.

Borovickova, Katarina, and Robert Shimer. 2017. "High Wage Workers Work for High Wage Firms." National Bureau of Economic Research NBER Working Paper 24074.

Brooks, Wyatt J., Joseph P. Kaboski, Yao Amber Li, and Wei Qian. 2019. "Exploitation of Labor? Classical Monopsony Power and Labor's Share." National Bureau of Economic Research NBER Working Paper 25660.

Brown, Charles, and James Medoff. 1989. "The employer size-wage effect." *Journal of Political Economy*, 97(5): 1027–1059.

Burdett, K., and D.T. Mortensen. 1998. "Wage differentials, employer size, and unemployment." *International Economic Review*, 39(2): 257–73.

Bureau of the Census. 2016*a*. "Longitudinal Business Database 1976-2014 [Data file]."

Bureau of the Census. 2016*b*. "Standard Statistical Establishment Listing 1976-2013 [Data file]."

Card, David, Ana Rute Cardoso, Jörg Heining, and Patrick Kline. 2018. "Firms and labor market inequality: Evidence and some theory." *Journal of Labor Economics*, 36(S1): 13–70.

Chan, Mons, Sergio Salgado, and Ming Xu. 2020. "Heterogeneous Passthrough from TFP to Wages." *Available at SSRN 3538503*.

Delgado, Mercedes, Richard Bryden, and Samantha Zyontz. 2014. "Categorization of traded and local industries in the US economy." Mimeo. Available online at: http://www. clustermapping. us/[Accessed May 2015].

De Loecker, Jan, Jan Eeckhout, and Gabriel Unger. 2020. "The Rise of Market Power and the Macroeconomic Implications." *The Quarterly Journal of Economics*, 135(2): 561–644.

Derenoncourt, Ellora, Clemens Noelke, and David Weil. 2021. "Spillover Effects from Voluntary Employer Minimum Wages." *Working Paper*.

Dube, Arindrajit, Jeff Jacobs, Suresh Naidu, and Siddharth Suri. 2020. "Monopsony in Online Labor Markets." *American Economic Review: Insights*, 2(1): 33–46.

Edmond, Chris, Virgiliu Midrigan, and Daniel Yi Xu. 2018. "How Costly Are Markups?" National Bureau of Economic Research NBER Working Paper 24800.

Friedrich, Benjamin, Lisa Laun, Costas Meghir, and Luigi Pistaferri. 2019. "Earnings Dynamics and Firm-Level Shocks." National Bureau of Economic Research NBER Working Paper 25786.

Giroud, Xavier, and Joshua Rauh. 2019. "State Taxation and the Reallocation of Business Activity: Evidence from Establishment-Level Data." *Journal of Political Economy*, 127(3).

Greenwood, Jeremy, Zvi Hercowitz, and Gregory W Huffman. 1988. "Investment, Capacity Utilization, and the Real Business Cycle." *American Economic Review*, 78(3): 402–417.

Gutiérrez, Germán, and Thomas Philippon. 2017. "Declining Competition and Investment in the U.S." National Bureau of Economic Research Working Paper 23583.

Handwerker, Elizabeth Weber, and Matthew Dey. 2018. "Megafirms and Monopsonists: Not the same employers, not the same workers." *BLS Working Paper Series*, 2.

Hershbein, Brad, Claudia Macaluso, and Chen Yeh. 2020. "Monopsony in the U.S. labor market." Working paper.

Hopenhayn, Hugo A. 1992. "Entry, Exit, and firm Dynamics in Long Run Equilibrium." *Econometrica*, 60(5): 1127–1150.

Hotelling, Harold. 1990. "Stability in competition." In *The Collected Economics Articles of Harold Hotelling*. 50–63. Springer.

Jarosch, Gregor, Jan Sebastian Nimcsik, and Isaac Sorkin. 2019. "Granular Search, Market Structure, and Wages." National Bureau of Economic Research NBER Working Paper 26239.

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

Kehrig, Matthias, and Nicholas Vincent. 2021. "The Micro-Level Anatomy of the Labor Share Decline." *The Quarterly Journal of Economics*, 136(2): 1031–1087.

Kennan, John, and James R Walker. 2011. "The effect of expected income on individual migration decisions." *Econometrica*, 79(1): 211–251.

Kline, Patrick, Neviana Petkova, Heidi Williams, and Owen Zidar. 2019. "Who Profits from Patents? Rent-Sharing at Innovative Firms." *Quarterly Journal of Economics*, 134(3): 1343–404.

Kroft, Kory, Yao Luo, Magne Mogstad, and Bradley Setzler. 2020. "Imperfect Competition and Rents in Labor and Product Markets: The Case of the Construction Industry." National Bureau of Economic Research.

Lallemand, Thierry, Robert Plasman, and François Rycx. 2007. "The Establishment-size Wage Premium: Evidence from European Countries." *Empirica*, 34(5): 427–451.

Lamadon, Thibaut, Magne Mogstad, and Bradley Setzler. 2019. "Imperfect Competition and Rent Sharing in the U.S. Labor Market." National Bureau of Economic Research NBER Working Paper 25954.

Loecker, Jan De, Jan Eeckhout, and Simon Mongey. 2021. "Quantifying Market Power and Business Dynamism in the Macroeconomy." National Bureau of Economic Research NBER Working Paper 28761.

MacKenzie, Gaelan. 2019. "Trade and market power in product and labor markets." *Unpublished Working Paper*.

Manning, Alan. 2003. *Monopsony in motion: Imperfect competition in labor markets.* Princeton University Press.

Manning, Alan. 2011. "Imperfect Competition in the Labor Market." In *Handbook of labor economics*. Vol. 4, 973–1041. Elsevier.

Rinz, Kevin. 2018. "Labor Market Concentration, Earnings Inequality, and Earnings Mobility." Center for Administrative Records Research and Applications , US Census Bureau CARRA Working Paper Series 2018-10.

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

Rossi-Hansberg, Esteban, Pierre-Daniel Sarte, and Nicholas Trachter. 2018. "Diverging Trends in National and Local Concentration." National Bureau of Economic Research NBER Working Paper 25066.

Salop, Steven C. 1979. "Monopolistic Competition with Outside Goods." *The Bell Journal of Economics*, 141–156.

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): 211–236.

Suárez Serrato, Juan Carlos, and Owen Zidar. 2016. "Who Benefits from State Corporate Tax Cuts? A Local Labor Markets Approach with Heterogeneous Firms." *American Economic Review*, 106(9): 2582–2624.

Verboven, Frank. 1996. "The nested logit model and representative consumer theory." *Economics Letters*, 50(1): 57–63.

Webber, Douglas A. 2015. "Firm market power and the earnings distribution." *Labour Economics*, 35: 123–134.

Appendix For Print Publication

A Efficient allocation

Planner problem. To measure and decompose the welfare losses due to labor market power we define an efficient benchmark. The planner's problem is to choose employment at all firms $\{n_{ijt}\}$ and capital K_t to maximize the present discounted value of utility subject to the definitions of preferences and technology and the aggregate resource constraint:

$$\mathbf{C}_t + \left[K_{t+1} - (1-\delta) K_t\right] = \int_0^1 \sum_{i=1}^{m_j} \overline{Z} z_{ijt} \left(k_{ijt}^{\gamma} n_{ijt}^{1-\gamma}\right)^{\alpha} dj \tag{A1}$$

The efficient allocation is characterized by the following first order condition for n_{ijt}:

$$-U_N(\mathbf{C}_t, \mathbf{N}_t) \left(\frac{n_{ijt}}{\mathbf{N}_{jt}}\right)^{\frac{1}{\eta}} \left(\frac{\mathbf{N}_{jt}}{\mathbf{N}_t}\right)^{\frac{1}{\theta}} = U_C(\mathbf{C}_t, \mathbf{N}_t)\ mpl_{ijt} \quad , \quad mpl_{ijt} = \alpha\gamma \frac{y_{ijt}}{n_{ijt}} \quad , \quad \text{for all} \quad ij \tag{A2}$$

On the right is the marginal product of labor at firm ij, converted into utils, while on the left is the disutility of supplying that labor transformed into utils. The marginal product of capital is equated across firms.[58] In this economy the aggregate markdown is $\boldsymbol{\mu}^* = 1$ and misallocation $\boldsymbol{\Omega}^* = 1$.

In a competitive equilibrium the allocation associated with the efficient allocation can be obtained if firms take their wage w_{ijt} as given. In this case equation (A2) corresponds to the firm's first order condition for n_{ijt}, combined with the household's aggregate labor supply curve. The wages that would be obtained in this case obviously correspond to the shadow wages of the planner, as such we use them to compute objects like the HHI_t^{wn} implied by the efficient allocation. This also justifies our description of the efficient allocation having *more competition* than the benchmark economy, since in the corresponding decentralization firms are competitive, taking their wages as given.

[58] First order condition for capital k_{ijt} equates the marginal product of capital at all firms to the shadow value of capital R_t^* which satisfies $U_C(\mathbf{C}_t, \mathbf{N}_t) = \beta U_C(\mathbf{C}_{t+1}, \mathbf{N}_{t+1})\left[R_t^* + (1-\delta)\right]$.