Unbundling Market Power[*]

Fabian Trottner
UC, San Diego

April 2023

How do the micro origins of firms' market power impact its aggregate implications? To answer this question, I develop a model of imperfect competition in output and factor markets with fixed costs; elastic factor supply; and variable markups and markdowns. I solve the planner's problem to characterize distortions, and show that the equilibrium behavior of firms and aggregates can be characterized in terms of sufficient statistics capturing departures from an observationally equivalent benchmark economy with competitive factor markets. I apply my results to decompose the societal costs of market power into three effects: (1) direct losses from inefficiencies in relative firm sizes, entry, and exit, (2) a deadweight loss, and (3) indirect efficiency gains or losses resulting from the interplay of competition and factor supply. Unbundling firms' market power, conditional on a set of common statistics, is not required for evaluating the first two effects; however, it is crucial for assessing indirect efficiency losses quantitatively and qualitatively. A quantitative exercise illustrates the importance of these results: indirect efficiency losses double and entry subsidies are much more effective once the interplay of markups and markdowns is accounted for.

[*] contact: ftrottner@ucsd.edu. I thank, without implicating, Fabian Eckert, Kyle Herkenhoff, Pablo Kurlat, Monica Morlacco, Marc Muendler, Romain Ranciere, and seminar participants at the Minneapolis Fed, USC, UC Irvine, UCLA, Queen's University, and SITE for thoughtful comments and suggestions. This version supersedes a previous draft circulated under the title "Misallocations in Monopsonistic Labor Markets."

1 Introduction

Firms' market power is a key determinant of welfare in product and factor markets. By compensating firms for sunk entry costs, it benefits consumers by enhancing product variety and those that supply the factors of production by channeling aggregate returns to scale. Yet, when exploited, it can cause factor misallocations and aggregate productivity losses. However, the role of product and factor markets in catalyzing these costs and benefits remains nebulous, partly because macroeconomic models routinely assume either of these markets to be perfectly competitive. Meanwhile, mounting evidence indicates that firms simultaneously exercise market power when competing for factors - e.g., labor - and costumers, suggesting the need for a unified approach to assessing the consequences of market power.

This paper delineates the implications of product and factor market power in a tractable model of imperfect competition featuring fixed costs, elastic aggregate factor supply, and variable markups in product and markdowns in factor markets. I analytically show that the equilibrium behavior of firms and aggregates can be characterized in terms of sufficient statistics that capture departures from an observationally equivalent benchmark economy with competitive factor markets. Applying this insight, I show that the welfare loss from market power can be decomposed into three effects. The first captures the direct costs of allocative inefficiencies in relative firm sizes, entry and exit, and holding fixed factor endowments. The second effect captures a deadweight loss that arises when factor supply is elastic. The final channel describes how interactions between allocative efficiency and factor supply shape welfare. In a quantitative illustration, I show that abstracting from markdowns not only understates overall welfare losses, but also the importance of the indirect efficiency effect. Indirect efficiency losses are substantially more important once markdowns are accounted for, implying that subsidies to entry are a viable tool to counter the costs of market power.

Section 2 describes the model. Households incur costs to supply and allocate the only factor of production - labor - across producers. Firms compete for factors, i.e., workers, and customers, by setting wages and output prices. Market power arises as jobs are differentiated in the returns they generate for households. I utilize a new factor supply system that implies even atomistic firms face upward-sloping labor supply curves with variable elasticities; moreover, it is homothetic, imposes no restrictions on how wage elasticities vary with firm size, and nests benchmark models of static monopsony.[1] To model product demand, I use a generalization of Kimball (1995) preferences introduced by Matsuyama & Ushchev (2017). Firms may face differently shaped residual product demand and labor supply curves and differ in productivity. Optimal

[1] See, e.g. Manning (2021) and Card *et al.* (2018). In the main text, I provide a microfoundation that illustrates how the model relates to this literature.

wages are a markdown relative to the marginal product of labor, and prices equal a markup over marginal costs. Markdowns and markups vary endogenously across firms of different types.[2]In equilibrium, prices and allocations are jointly determined with firm entry and exit decisions and factor supply decisions of households.

Section 3 solves the planner's problem to identify distortions in the economy. I show that when labor supply is inelastic, i.e., primary factors are in fixed supply, the decentralized allocation is efficient if, and only if, markdowns and markups are homogeneous across firms. This uniquely ensures that the rents[3] earned by each producer coincide with the combined surpluses it generates for workers and consumers. As a result, the market internalizes all relevant externalities and induces socially optimal allocations. In contrast, when markdowns or markups differ across producers, the market allocation is inefficient due to relative firm sizes, entry, and selection distortions. I provide statistics that characterize each source of inefficiency, highlighting that effective markups - the bundled price markup to wage markdown ratio - characterize distortions of both the entry and relative firm size margin. When aggregate labor supply is elastic, the decentralized equilibrium is inefficient. Intuitively, markdowns and markups jointly act like a uniform tax on consumption or, alternatively, a labor wedge.[4] This lowers labor supply compared to the first-best. The implied deadweight loss reduces welfare even if factor allocations are efficient, my analysis delineates how its magnitude is, in general equilibrium, initimately intertwined with realized allocative inefficiencies.

Section 4 characterizes the general equilibrium response of total factor productivity to changes in market size. Market size may change endogenously via factor supply, or exogenously due to, say, policy intervention,or change in population. I derive a set of sufficient statistics, which I refer to as effective price and cost pass-throughs, that summarize relevant details of a firm's underlying market power to characterize its exposure to price competition. Conceptually, these statistics capture the extend to which firm behavior in the model is isomorphic to workhorse models of variable markups with competitive factor markets.[5] A second set of statistics informs deviations in firm

[2] The model imposes no particular relationship between, say, a firm's employment relative to the market and its markups or markdowns. Further, my theoretical results easily extend to an economy where overhead costs vary across firms and differences in profit-shares, hence, partly reflect fixed costs rather than market power.

[3] Rents are measured in terms of the surplus derived from an infra-marginal job or variety. Factor rents correspond to the excess return over that required to change the choice of employer, as in Robinson (1933), Rosen (1987), and, more recently, Lamadon *et al.* (2022). Firm rents correspond to effective markups.

[4] This point goes back as early as Lerner (1934) and Samuelson (1947).

[5] As a key implication, existing estimates of price cost pass-throughs by, e.g. ,Amiti *et al.* (2019) continue to inform key elastcities in a model with imperfect competition in factor and product markets. Consequently, cost pass-throughs are also sufficient to characterize wage pass-throughs if the product market is competitive.

behavior from this benchmark as interactions between markups and markdowns leave firms differentially exposed to competitive pressures in product and factor markets.

Leveraging this insight, I characterize when unbundling the source of firms' market power is important for describing the equilibrium responses of firms and aggregate outcomes to shocks. In the absence of distortions, the source of firms' markups is irrelevant for policy and welfare analysis. When market power originates from only one market, then conditional on a set of effective markups, pass-throughs, and demand elasticities, factor and product market power imply the same firm-level responses to a given set of shocks, and differences between economies with factor market power are solely due to the fact that factors employed in fixed cost production earn rents. When effective markups are insufficient to characterize the equilibrium response to shocks, new reallocative channels emerge, which may counter or amplify, e.g., the pro-competitive effects of market size. My results show how these effects intuitively depend on the micro-level distribution of markups and markdowns, and how their quantitative importance is summarized by a few aggregate statistics.

Section 5 characterizes the welfare change from optimal policies, which coincide with the social costs of these distortions. Up to a second-order approximation, the distance to the efficient frontier can be analyzed in terms of how productivity and welfare respond to removing distortions. I show that the macroeconomic effects of micro-level heterogeneity in markdowns and markups can be characterized through three effects. First, the direct efficiency effect corresponds to the change in welfare that can be achieved by removing inefficiencies in entry, selection, and variable factor allocations, while holding the total supply of factors fixed. The analytical expression neatly decomposes this effect into the contributions of each of these margins of inefficiency. Removing distortions in factor supply, like an increase in population, triggers changes in both technical and allocative efficiency. The former captures the welfare effects of an increase in factor supply when allocations are held fixed. The latter captures how adjustments in entry, markups, markdowns, and selection contribute to welfare. Finally, each margin is summarized by intuitive sufficient statistics showing that knowledge of firms' factor and product market power, conditional on effective markups, cost pass-throughs and demand elasticities, is not required to evaluate the weflare loss due to the direct and technical efficiency effect. In contrast, unbundling markups and markdowns is crucial to assessing indirect efficiency losses quantitatively and qualitatively..

Section 6 quantifies the theoretical results and traces policy implications. I develop a strategy that, given estimates of markdowns and markups across firms, allows me to recover the firm-level elasticities required to quantify and decompose the welfare loss in the economy. I implemnt this approach using estimates of markdowns and markups for German manufacturing firms by Dolfen (2020). The calibrated model displays substantial dispersion in markdowns and markups systematically related to sales and pay,[6] near-complete passthrough of cost and productivity shocks into prices and wages for small producers and substantially passthrough rates for both prices and wages for the largest firms.[7]

Quantitatively, I find that the costs of imperfect competition can be high, and that allocative inefficiencies account for about 70 percent of the overall welfare loss, suggesting that heterogeneity in markups and markdowns is critical for the costs of distortions. The endogenous interaction between factor supply and allocative efficiency greatly compounds allocative losses, nearly doubling the costs of those implied by the direct efficiency effect.As a key policy implication, subsidies to entry costs, by harnessing indirect efficiency gains, are an effective remedy to the inefficiencies caused by imperfect competition.

Finally, I assess the importance of accounting for factor market competition. To do so, I reassess welfare losses imposing that labor markets are perfectly competitive and observed markups are fully attributable to price markups, showing that welfare losses would be lower. Since the counterfactual holds the social costs of technical efficiencies fixed, this shows that factor market power, on the net, compound allocative inefficiencies posed by product market power. However, the composition of these costs can change dramatically., and reflect almost exclusively the direct efficiency effect. In part, this reflects changes in markups and reallocations triggered by entry jointly leave the aggregate deadweight loss largely unchanged. In contrast, insufficient competition in product markets partially inhibits the reallocations towards the largest firms, while also directly lowering the labor wedge by raising aggregate quasi-rents - the share of profits used to finance sunk entry costs.

Related literature. This paper relates closely to the literature studying the welfare costs of markups. Early contributions by Spence (1976), Dixit & Stiglitz (1977), Venables (1985), Mankiw & Whinston (1986), and recent work by Matsuyama & Ushchev (2020) analyse the welfare effects of variable markups in models with homogeneous firms. Zhelobodko *et al.* (2012), Dhingra & Morrow (2019), and Behrens *et al.* (2020) study static models with heterogeneous firms, while Bilbiie *et al.* (2019) analyze a dynamic model with homogeneous and Edmond *et al.* (2021) one with heterogeneous firms. I provide an integrated framework for studying the welfare implications of im-

[6] Following approaches consistent with the estimates of Dolfen (2020), Yeh *et al.* (2022) find substantial variation in markdowns and markups across U.S. manufacturing plants, and Brooks *et al.* (2021) find similar patterns for indian plants.

[7] For example, Amiti *et al.* (2019) provide evidence that price cost pass-throughs vary systematically across Belgian manufacturing firms by sales share. Chan *et al.* (2021) provide evidence that pass-throughs of productivity shocks into wages are higher for firms with high compared to those with low employment shares in Denmark.

perfect competition in both labor and product markets that nests many of the insights generated by this literature as a special case.

My work also relates to work in international trade studying the gains from market size in models with markups, such as Krugman (1979), Melitz (2003), Melitz & Ottaviano (2008), Epifani & Gancia (2011), Arkolakis *et al.* (2012), Melitz & Redding (2015), Mrázová & Neary (2017, 2019), Arkolakis *et al.* (2019), and Matsuyama & Ushchev (2022). Many of the theoretical results provided in this paper build on and are inspired by the body of recent work by Baqaee *et al.* (2022) who analyze the gains from an increase in market size in a model with variable markups. I contribute by providing the first assessment of how imperfect competition in factor markets interacts with the forces stressed by this literature.

Further, I contribute to the literature studying the implications of monopsony. Theoretically, I generalize a class of benchmark models of static monopsony, described in e.g., Manning (2003), Card *et al.* (2018), Trottner (2020), Haanwinckel (2021), Jha & Rodriguez-Lopez (2021), Kroft *et al.* (2020), and Lamadon *et al.* (2022), which utilizes logit discrete job choice models to generate firm-level labor supply curves with a constant wage elasticity. Building on Thisse & Ushchev (2016), I generalize this approach to rationalize homothetic, flexible labor supply systems with variable wage elasticities. Thereby, even a model with atomistic firms can accommodate key empirical features of markdowns and wage passthroughs documented in, e.g., Staiger *et al.* (2010), Webber (2015), Serrato & Zidar (2016), Garin & Silvero (2018), Chan *et al.* (2019), Dolfen (2020), Dube *et al.* (2020), and Yeh *et al.* (2022). Further, I show that alternative functional forms introduced to generalize CES demand by Matsuyama & Ushchev (2017) could equally well be used to account for variable markdowns, endogenous entry, and exit.

My work abstracts from strategic interactions in price- and wage-setting. In recent work, Berger *et al.* (2022) adapt the seminal approach to modeling variable markups under oligopsony by Atkeson & Burstein (2008) to assess the costs of labor market power when product markets are competitive and entry is exogenous. Edmond *et al.* (2021) build on Atkeson & Burstein (2008) to assess the costs of markups under free entry, showing how assumptions on the market structure impact welfare analysis. While solving models with entry and exit, firm heterogeneity and strategic interactions in price- and wage-setting remains an unsolved problem, my results provide a first step toward establishing when models of product and factor markets with oligopolistc market structures may, in fact, be isomorphic to each other.

2 Theoretical Framework

This section lays out the problem of agents in the economy, and defines the decentralized equilibrium.

2.1 Description of the model

2.1.1 Households

Preferences The economy is populated by a mass L of identical households. Each household derives utility from consumption $\mathcal{Y}$ and disutility from supplying the only factor of production $\mathcal{N}$, which I henceforth refer to as labor.

$$\mathcal{U} = U(\mathcal{Y}, \mathcal{N}). \tag{1}$$

The utility index U is twice continously differentiable and satisfies standard properties.[8] Households consume varieties $\theta \in \Theta$ and allocate labor to jobs $\omega \in \Omega \equiv \Theta \bigcup \{e, o\}$.[9] The consumption index $\mathcal{Y}$ and labor supply index $\mathcal{N}$ are defined as:

$$1 = \int_{\Theta} \Upsilon_{\theta}(\frac{y_{\theta}}{\mathcal{Y}}) dM^{C}(\theta), \tag{2}$$

$$1 = \int_{\Omega} \Psi_{\omega}(\frac{n_{\omega}}{N}) dM^{E}(\omega), \tag{3}$$

where y_θ denotes per-capita consumption of a variety of type θ, and n_ω denotes per-capita labor allocations to an employer of type ω. The masses of varieties $dM^{\mathcal{Y}}(\theta)$ of type θ and jobs $dM^{E}(\omega)$ of type ω are described further below. The consumption utility indices $\Upsilon_\theta(.)$ are strictly increasing, concave, and satisfy $\Upsilon_\theta(0) = 0$. The labor cost indices $\Psi_\omega(.)$ are strictly increasing, convex, and satisfy $\Psi_\omega(0) = 0$.

Introduced by Matsuyama & Ushchev (2017), the consumption preferences in (2) generalize the Kimball (1995) demand system.[10] I use an analogous functional form to flexibly model households' preferences over factor allocations. CES preferences over consumption varieties and jobs are nested as special cases when $\Upsilon_\theta(x) = a_\theta x^{\frac{\sigma-1}{\sigma}}$ and $\Psi_\omega(x) = b_\omega x^{\frac{\beta+1}{\beta}}$, where a_θ and b_ω denote exogenous taste/efficiency shifters.

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

[9] The production structure in the economy is explained in more detail further below.

[10] Matsuyama & Ushchev (2017) refer to (2) as the class of homothetic demand systems with implicit additivity (HDIA). The autors also introduce the class of homothetic demand systems with a single aggregator (HSA) as an alternative way to generalize CES preferences. See Appendix D for a version of the model with HSA-type product demand and labor supply systems.

Utility Maximization Consumers maximize utility subject to the following budget constraint,

$$\int_{\Theta} p_\theta y_\theta dM^C(\theta) = \int_{\Omega} n_\omega w_\omega dM^E(\omega) = 1,$$

where p_θ is the price of a variety of type θ and w_ω is the wage offered by an employer of type ω. Total earnings are chosen as the numeraire.

The per-capita inverse demand for variety θ is given by,[11]

$$\frac{p_\theta}{\mathcal{P}} = \Upsilon'_\theta(\frac{y_\theta}{\mathcal{Y}}), \tag{4}$$

and per-capita factor supply to job ω equals,

$$\frac{w_\omega}{\mathcal{W}} = \Psi'_\omega(\frac{n_\omega}{\mathcal{N}}), \tag{5}$$

where $\mathcal{P}$ and $\mathcal{W}$ are price and wage aggregates given by,

$$\mathcal{P} \equiv \frac{\bar{P}}{\mathcal{Y}}, \qquad \frac{1}{\bar{P}} \equiv \int_{\Theta} \Upsilon'_\theta(\frac{y_\theta}{\mathcal{Y}})\frac{y_\theta}{\mathcal{Y}} dM^C(\theta), \tag{6}$$

$$\mathcal{W} \equiv \frac{\bar{W}}{\mathcal{N}}, \qquad \frac{1}{\bar{W}} \equiv \int_{\Omega} \Psi'_\omega(\frac{n_\omega}{\mathcal{N}})\frac{n_\omega}{\mathcal{N}} dM^E(\omega). \tag{7}$$

Following equations (4) and (5), the indices Ψ_θ and Υ_ω, respectively, determine the shape of the residual product demand and factor supply faced by firms. The demand for a variety depends on its price relative to the demand shifter $\mathcal{P}$, while factor supply to individual jobs depends on the offered wage relative to the aggregate $\mathcal{W}$, showing that the shifters $\mathcal{P}$ and $\mathcal{W}$ mediate competition between firms.[12]

To choose $\mathcal{Y}$ and $\mathcal{N}$, household's supply factors until the utility cost is equal to the real wage, which also denotes aggregate factor productivity $\mathcal{A}$, $-\frac{U_\mathcal{N}}{U_\mathcal{Y}} = \frac{\mathcal{Y}}{\mathcal{N}} = \mathcal{A}$.

Microfoundation As will be discussed momentarily, producers in the model are atomistic but have endogenously varying degrees of wage- and price-setting power. The factor supply system in (5),provides a novel and highly tractable model of endogenous competition in broadly defined factor markets. Appendix A.2 develops a microfoundation based on a model of random discrete choice. Building on Thisse &

[11] The derivation is relegated to Appendix A.1.

[12] In general, the real wage is not equal to $\mathcal{W}/\mathcal{P}$. Letting $\mathcal{P}^I$ and $\mathcal{W}^I$ denote the wage indices that solve the nested expenditure minimization and income maximization problems, $e(\{p_\omega\}, C) = \mathcal{P}^I C$, $I(\{w_{\omega'}\}, N) = \mathcal{W}^I N$, the household maximizes $\mathcal{U}(C, N)$ subject to $\mathcal{P}^I C = \mathcal{W}^I N$. By definition, $d\log\frac{C}{N} = d\log\mathcal{W}^I - d\log\mathcal{P}^I$, while $d\log\frac{\mathcal{W}}{\mathcal{P}} = d\log\frac{C}{N} + d\log\bar{W} - d\log\bar{P}$. These expressions coincide if, and only if, labor supply and product demand systems are of the CES type, that is $\frac{\partial\log\Psi_\omega(x)}{\partial\log x} \equiv \frac{\beta+1}{\beta}$ and $\frac{\partial\log\Upsilon_\omega(x)}{\partial\log x} \equiv \frac{\sigma-1}{\sigma}$.

Ushchev (2016), I show that by allowing indirect utility comparisons between jobs to depend on the entire set of choices, one can rationalize a rich class of factor supply systems with variable elasticities. Job differentiation may arise from job-specific taste or productivity shocks, highlighting that the factor supply system in (5) lends itself to model imperfect competition in settings where market power arises from broadly defined information frictions. [13]

2.1.2 Firms

Variety Producers Final good firms produce differentiated varieties and compete monopolistically in product and monopsonistically in labor markets. To enter, a prospective producer must obtain a fixed quantity f_e of entry inputs at per-unit cost p_e. Upon entry, final good producers receive a type draw θ from a continuous distribution with cdf $G(\theta)$, density $g(\theta)$, and compact support supp$(G) \subset \mathbb{R}$. After receiving its draw θ, a firm decides whether to produce or exit. Production requires obtaining fixed quantity f_o overhead inputs o at price p_o, and a firm of type θ produces A_θ units of output per unit of employed labor. Firms decide which price and wage to set, taking as given their labor supply curve (5), product demand curve (4), and productivity A_θ.

From (4), the price elasticity of demand faced by a firm of type θ is given by,

$$\sigma_\theta(\frac{y}{\mathcal{Y}}) \equiv -\frac{\partial \log y_\theta}{\partial \log p_\theta} = -\frac{\Upsilon'_\theta(\frac{y}{\mathcal{Y}})}{\Upsilon''_\theta(\frac{y}{\mathcal{Y}})\frac{y}{\mathcal{Y}}}, \tag{8}$$

and, following (5), the wage elasticity of its factor supply equals,

$$\beta_\theta(\frac{n}{\mathcal{N}}) \equiv \frac{\partial \log n_\theta}{\partial \log w_\theta} = \frac{\Psi'_\theta(\frac{n}{\mathcal{N}})}{\Psi''_\theta(\frac{n}{\mathcal{N}})\frac{n}{\mathcal{N}}}. \tag{9}$$

Conditional on operating, the desired price of a firm of type θ is a markup μ_θ over its marginal cost mc_θ:[14]

$$p_\theta = \mu_\theta(\frac{y_\theta}{\mathcal{Y}})mc_\theta,$$

where the markup is given by,

$$\mu_\theta(\frac{y_\theta}{\mathcal{Y}}) = \frac{\sigma_\theta(\frac{y_\theta}{\mathcal{Y}})}{\sigma_\theta(\frac{y_\theta}{\mathcal{Y}}) - 1}, \tag{10}$$

Profit-maximizing wages, in turn, equal a markdown $\mathcal{M}_\theta$ over a firm's marginal rev-

[13] Through this lense, the disutility from supplying a total amount of labor N in preferences captures the opportunity cost of factor production. For example, if N encapsulated physical capital instead of labor, it would capture the steady state costs of transforming some generic endowment into consumption.

[14] I assume Υ_θ and Ψ_θ are such that marginal profits are strictly decreasing for all $\theta \in \text{supp}G$.

enue product of labor $mrpl_\theta$:

$$w_\theta = \mathcal{M}_\theta(\frac{n_\theta}{\mathcal{N}})mrpl_\theta,$$

and the desired markdowns depend inversely on the wage elasticity of labor supply,

$$\mathcal{M}_\theta(\frac{n_\theta}{\mathcal{N}}) = \frac{\beta_\theta(\frac{n_\theta}{\mathcal{N}})}{\beta_\theta(\frac{n_\theta}{\mathcal{N}})+1}. \tag{11}$$

Wages and prices, thus, are related through the following,

$$p_\theta = \frac{\mu_\theta}{\mathcal{M}_\theta}\frac{w_\theta}{A_\theta} = \boldsymbol{\mu}^e_\theta\frac{w_\theta}{A_\theta}, \tag{12}$$

where $\boldsymbol{\mu}_\theta \equiv \frac{\mu_\theta}{\mathcal{M}_\theta}$ indicates a firm's effective markup. Together, (4), (5), and (12) determine a firm's price, wage and labor demand. Depending primitives, markups and markdowns may be homogeneous and constant across firms, vary exogenously with a firm's type θ, or vary endogenously with a firm's relative employment and output. E.g., in the special case where firms face labor supply curves with constant wage elasticities β_θ, $\Psi_\theta(x) = x^{\frac{\beta_\theta+1}{\beta_\theta}}$, markdowns vary exogenously by firm type if $\beta_\theta \neq \beta_{\theta'}$, and are constant across firm types if $\beta_\theta = \beta$. Away from these special cases, markdowns and markups also vary endogenously with firms' relative employment n_θ/N and relative output $y_\theta/\mathcal{Y}$.

A firm operates if, and only if, its variable profits exceed its overhead costs:

$$Lp_\theta c_\theta\left(1 - \frac{\mathcal{M}_\theta}{\mu_\theta}\right) \geq p_o f_o. \tag{13}$$

Assuming that firm types are ordered so that operating profits are strictly increasing and continuously differentiable in θ,[15] there exists a unique cutoff θ^* such that firms with types $\theta \geq \theta^*$ produce, while firms with types $\theta < \theta^*$ exit the market.

Free entry implies that expected operating profits upon entry are equal to the entry cost:

$$\int_{\theta\geq\theta^*}\left(L(1 - \frac{\mathcal{M}_\theta}{\mu_\theta})p_\theta y_\theta - p_o f_o\right)dG\left(\theta\right) = p_e f_e. \tag{14}$$

The measure of a firms of type θ is given by $dM^C(\theta) = dM^E(\theta) = Mg(\theta)\mathbf{1}_{\{\theta>\theta^*\}}d\theta$, where M is the mass of entrants.

Entry and Overhead Inputs Entry and overhead inputs are indivisible, and produced by homogeneous firms using a linear production technologies. The market

[15] In terms of primitives, this requires

for overhead and entry inputs is perfectly competitive. Producers enter freely and compete for workers in the nationwide labor market.[16] Given these assumptions, each entry input is provided by a single producer at a price p_e that equals the average cost of hiring a total f_e hours of labor,

$$p_e = \mathcal{W}\Psi_e'(\frac{f_e}{L\mathcal{N}}). \tag{15}$$

and the mass of entry input producers equals the mass of entrants, $dM^E(e) = M$. Similarly, the mass of overhead input producers equals $dM^E(o) = M[1 - G(\theta^*)]$ and the overhead unit price p_o equals,

$$p_o = \mathcal{W}\Psi_o'(\frac{f_o}{L\mathcal{N}}). \tag{16}$$

2.1.3 Equilibrium

An equilibrium consists of a mass of entrants M, an exit cutoff θ^*, allocations $\{n_o, n_e, \{c_\theta, n_\theta\}_{\theta\in\Theta}\}$, and prices $\{p_o, p_e, \{p_\theta, w_\theta\}_\theta\}$ such that consumers maximize utility taking prices and wages as given, producers maximize profits, taking $\mathcal{Y}, \mathcal{N}, \mathcal{P}$ and $\mathcal{W}$ as given, and markets clear.

2.2 Notation

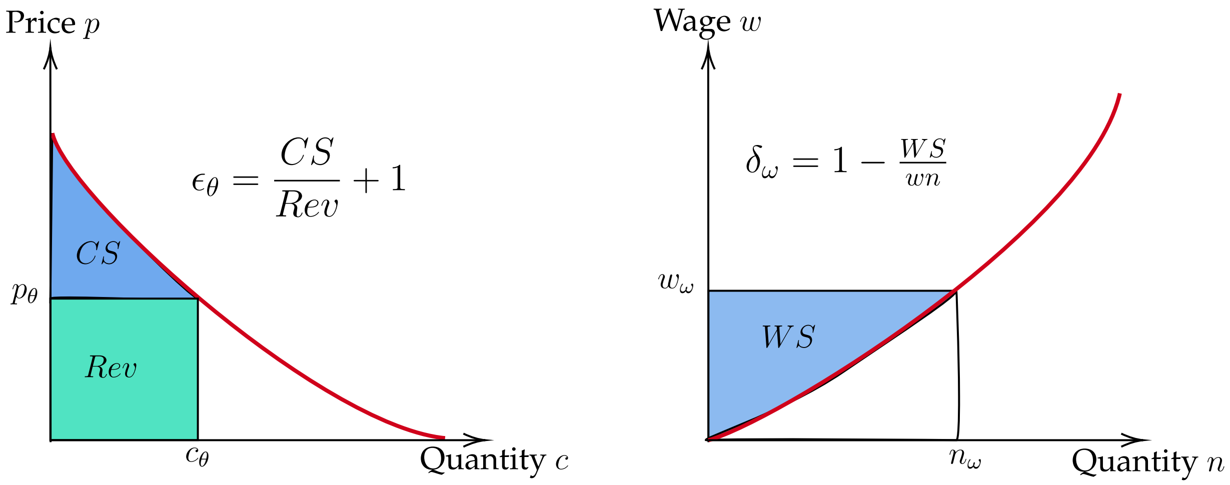
Notation Throughout the rest of this paper, the expectation of a variable x with respect to the sales density $s_\theta = \frac{p_\theta c_\theta}{\int_{\theta*}^{\infty} p_{\theta'} c_{\theta'} dG(\theta')}$ by $\mathbb{E}_s[x_\theta] \equiv \int_{\theta*}^{\infty} s_\theta x_\theta dG(\theta)$, and its wage-bill-weighted by $\mathbb{E}_{wn}[x_\omega] = \frac{w_e f_e x_e + \int_{\theta*}^{\infty}[w_o f_o x_o + w_\theta n_\theta x_\theta] dG(\theta)}{f_e w_e + \int_{\theta*}^{\infty}[w_o f_o + w_\theta n_\theta] dG(\theta)}$. The covariance of two variables x and z with respect to the sales density is denoted $Cov_s[x_\theta, z_\theta] = \mathbb{E}_s[x_\theta y_\theta] - \mathbb{E}_s[z_\theta]\mathbb{E}_s[y_\theta]$.

Household rents The social value of a job of type ω is summarized by $1 - \delta_\omega$, where δ_ω is the inframarginal disutility associated with supplying labor to ω,

$$\delta_\omega(\frac{n_\omega}{\mathcal{N}}) \equiv \frac{\Psi(\frac{n_\theta}{\mathcal{N}})}{\Psi'(\frac{n_\theta}{\mathcal{N}})\frac{n_\theta}{\mathcal{N}}} \in (0, 1]. \tag{17}$$

[16] Appendix A.3 develops an extension where overhead inputs are produced within and overhead requirements vary across firms. The assumption of homogeeous overhead input requirements is not crucial to the theoretical results. What is crucial, however, is that overhead and variable production jobs are differentiated from the perspective of the household, and that firms cannot internally move factors between uses.

Figure 1 Consumption and Employment surpluses



The right panel in Figure 1 shows that δ_ω equals the fractions of earnings that compensates workers for the utility costs of supplying labor. In a percetly competitive market, wages fully compensate workers for the disutility from work and firms capture the entire worker surplus, implying that $\delta_\omega = 1$. In general, the surplus a job of type ω generates a - rents in the sense of Rosen (1987) - equal to $(1 - \delta_\omega) w_\omega n_\omega$.[17]

Analogously, the infra-marginal consumption surplus ϵ_θ measures the value of a firm of type θ to consumers.

$$\epsilon_\theta(\frac{y_\theta}{\mathcal{Y}}) = \frac{\Upsilon_\theta(\frac{y_\theta}{\mathcal{Y}})}{\Upsilon'_\theta(\frac{y_\theta}{\mathcal{Y}})\frac{y_\theta}{\mathcal{Y}}} \geq 1. \tag{18}$$

The left panel in Figure 1 shows that ϵ_θ corresponds to 1 plus the ratio of consumer surplus to revenues, and firms of type θ generate a surplus for consumers equal to $(\epsilon_\theta - 1) p_\theta c_\theta$.

In general, consumption and worker surpluses differ endogenously across producers. For the remainder of the paper, let $\bar{\epsilon} = \mathbb{E}_s[\epsilon_\theta]$ and $\bar{\delta}$=$\mathbb{E}_{wn}[\delta]$ denote the sales- and wage-bill-weighted averages of infra-marginal surpluses in product and labor markets.[18]Again, I use a bold symbol to indicate households' effective surplus $\bar{\boldsymbol{\epsilon}} = \bar{\epsilon}/\bar{\delta}$.

Pass-Throughs The productivity wage pass-through γ_θ is the elasticity of firm-level wages with respect to shocks to its marginal revenue product of labor,

$$\gamma_\theta(\frac{w}{\mathcal{W}}) \equiv \frac{\partial \log w_\theta}{\partial \log mrpl_\theta} = \frac{1}{1 - \frac{\frac{w}{\mathcal{W}}\mathcal{M}_\theta'(\frac{w}{\mathcal{W}})}{\mathcal{M}_\theta(\frac{w}{\mathcal{W}})}}. \tag{19}$$

[17] The microfoundation for the labor supply system permits the interpretation that rents arise from information frictions.

[18] Note that $\bar{\epsilon}$ equals demand index $\bar{P}$ defined in (6), while $\bar{\delta}$ equals the labor supply index $\bar{W}$ defined in equation (7).

If labor supply is isoelastic, then shocks to a firms' labor productivity are fully passed on to wages, $\gamma_\theta = 1$. In general, markdowns and the pass-through γ_θ vary with offered wages. Similarly, a firm's price pass-through ρ_θ captures the elasticity of a firms' markup to its price, and is defined to satisfy,

$$\rho_\theta(\frac{p}{\mathcal{P}}) \equiv \frac{\partial \log p_\theta}{\partial \log mc_\theta} = \frac{1}{1 - \frac{\frac{p}{\mathcal{P}}\mu'_\theta(\frac{p}{\mathcal{P}})}{\mu_\theta(\frac{p}{\mathcal{P}})}}. \tag{20}$$

3 Efficiency

This section solves the problem of a social planner to characterize distortions in the decentralized economy.

3.1 The planner's problem

The planner's objective is to maximize per-capita welfare subject to the economy's technological constraints, i.e., the entry process and production technologies. Fixed costs imply that the planner chooses zero quantities for firms below a type threshold θ^*. Therefore, the planner's problem is given by,

$$\max_{y_\theta, n_\theta, \theta^*, M, n_o, n_e} U(\mathcal{Y}, \mathcal{N}),$$

subject to preferences, technological constraints, and the entry process,

$$1 = M \int_{\theta \geq \theta^*} \Upsilon_\theta(\frac{y_\theta}{\mathcal{Y}}) dG(\theta),$$

$$1 = M \left\{ \Psi_e(n_e/\mathcal{N}) + \int_{\theta \geq \theta^*}^{\infty} \left\{ \Psi_o(n_o/\mathcal{N}) + \Psi_\theta(\frac{n_\theta}{\mathcal{N}}) \right\} dG(\theta) \right\},$$

$$y_\theta \leq n_\theta A_\theta, L n_e \geq f_e, L n_o \geq f_o$$

A decentralized equilibrium is said to be efficient if it coincides with the allocation chosen by the planner.

3.2 Allocative Distortions

To isolate micro-level distortions, I begin by analyzing the planner's problem in an economy with inelastic aggregate factor supply. The following provides necessary and sufficient conditions for the market to efficiently allocate ressources across firms.[19] All

[19] See Section 4 for a discussion of equilibrium existence and uniqueness.

proofs are relegated to Appendix B.

Theorem 1. *Suppose factor supply is inelastic. The decentralized equilibrium is efficient if, and only if,* $\Upsilon_\theta(x) = a_\theta x^{\frac{\sigma-1}{\sigma}}$ *and* $\Psi_\omega(x) = b_\omega x^{\frac{\beta+1}{\beta}}$*, where* $\sigma, \beta > 1$*,* a_θ*,*$b_\omega \in \mathbb{R}^+$*.*

Theorem 1 shows that homogeneity in markdowns is necessary for the market allocation to be socially optimal. If, in addition, firms also have homogeneous degrees of market power in the product market, the market efficiently allocates of resources across productive uses. Thus, firm market power in neither input nor output markets is not per se a source of distortions, reflecting that profits are necessary to preserve entry incentives for producers in the presence of fixed costs. Thereby, Theorem 1 generalizes important insights from the literature on the welfare effects of monopolistic competition with heterogeneous firms (e.g., Zhelobodko *et al.* (2012), Dhingra & Morrow (2019), Edmond *et al.* (2021)) to an economy with imperfect competition in product and factor markets.

To gain intuition for this result, it is useful to compare the private and social surplus generated by each producer. Seeking to allocate labor so as to maximize each variety's "social profits" $\Upsilon_\theta(y_\theta^{\text{opt}}/\mathcal{Y}^{\text{opt}}) - \Psi_\theta(n_\theta^{\text{opt}}/\bar{\mathcal{N}})$, the planner prices goods at a "social markup" $\epsilon_\theta/\delta_\theta$,

$$\Upsilon_\theta(\frac{y_\theta^{\text{opt}}}{\mathcal{Y}^{\text{opt}}}) = \frac{\epsilon_\theta}{\delta_\theta}\Psi_\theta(\frac{n_\theta^{\text{opt}}}{\bar{N}}). \tag{21}$$

Private firms of type θ, in turn, choose quantities so as to maximize private profits,[20]

$$\Upsilon_\theta(\frac{y_\theta^{\text{mkt}}}{\mathcal{Y}^{\text{mkt}}}) = \frac{\mu_\theta}{\mathcal{M}_\theta}\Psi_\theta(\frac{n_\theta^{\text{mkt}}}{\mathcal{N}})\frac{\epsilon_\theta \mathbb{E}_{wn}\left[\delta_\omega\right]}{\delta_\theta \mathbb{E}_s[\epsilon_\theta]}. \tag{22}$$

When markdowns and markups are homogeneous, private production incentives, captured by $\mu_\theta/\mathcal{M}_\theta$, exactly coincide with social production incentives, $\epsilon_\theta/\delta_\theta$, which enables the market to incentivize exactly the right firms to produce. The proof formalizes this intuition, showing that the alignment of private and social production incentives at the firm-level implies efficient entry and selection. When entry and exit are efficient, competition in product and labor markets aligns the price and wage indices to ensure optimal firm-level quantities.

In contrast, heterogeneity in either markups or markdowns results in an allocation that inefficiently distributes the factors of production across producers. To characterize the resulting distortions in labor allocations across variety, entry, and overhead production, the following illustrates how feasible reallocations along each of these margins

[20] This follows from the fact that $p_\theta = \frac{\mu_\theta}{\mathcal{M}_\theta}\frac{w_\theta}{A_\theta}$ can be written as $C\mathcal{P}\frac{1}{\epsilon_\theta}\Upsilon(\frac{c_\theta}{C}) = \frac{\mu_\theta}{\mathcal{M}_\theta}N\mathcal{W}\frac{1}{\delta_\theta}\Psi(\frac{n_\theta}{N})$, and observing that $C\mathcal{P} = \mathbb{E}_{pc}\left[\epsilon_\theta\right]$ and $N\mathcal{W} = \mathbb{E}_{wn}[\delta]$.

impact welfare.[21]

Factor misallocation across entrants Consider a feasible reallocation of workers from firms in $(\theta', \theta' + d\theta')$ to those in $(\theta, \theta + d\theta')$. If this reallocation raises welfare, firm θ is said to be too large compared to θ'.

Lemma 1. *Reallocating variable labor from θ' to θ raises welfare if, and only if,*

$$\boldsymbol{\mu}^e_\theta \equiv \frac{\mu_\theta}{\mathcal{M}_\theta} > \frac{\mu_{\theta'}}{\mathcal{M}_{\theta'}} = \boldsymbol{\mu}^e_{\theta'}. \tag{23}$$

Lemma 1 shows that cross-sectional distortions factor allocations can be summarized by effective markups $\boldsymbol{\mu}_\theta$. In contrast to standard results in the literature on monopolistic competition,[22] firms with higher *price* markups μ_θ are not necessarily inefficiently small.However, knowledge of effective markups, while useful to inform distortions, is not sufficient to characterize the general equilibrium behavior of firms.

Entry distortion To assess distortions in entry, consider a reallocation that moves workers from variable to entry and overhead production, while leaving relative quantities $y_\theta/\mathcal{Y}$, selection, and aggregate labor supply unchanged. If such a reallocation raises welfare, entry is insufficient. Else, it is said to be excessive.

Lemma 2. *In a given allocation, entry is insufficient if, and only if,*

$$\bar{\boldsymbol{\epsilon}} \equiv {}^{\bar{\epsilon}}/_{\bar{\delta}} > \mathbb{E}_s\left[{}^{\mathcal{M}_\theta}/_{\mu_\theta}\right]^{-1} \equiv \mathbb{E}_s\left[{}^{1}/_{\boldsymbol{\mu}^e_\theta}\right]^{-1}. \tag{24}$$

where $\bar{\epsilon} = \mathbb{E}_s\left[\epsilon_\theta\right]$, $\bar{\delta} = \mathbb{E}_{wn}\left[\delta_\omega\right]$, $\bar{\boldsymbol{\epsilon}} = {}^{\bar{\epsilon}}/_{\bar{\delta}}$ and $\boldsymbol{\mu}^e_\theta \equiv \frac{\mu_\theta}{\mathcal{M}_\theta}$.

A marginal entrant, on average raises social profits, i.e., per-capita welfare, by $\frac{\bar{\epsilon}}{\bar{\delta}} - 1$, while reducing private profits by $\mathbb{E}_s\left[\mathcal{M}_\theta/\mu_\theta\right]^{-1} - 1$. Lemma 2 says that entry is insufficient if an increase in the the mass of entrants raises social profits by more than it reduces the average profits of firms. Consistent with Theorem 1, when effective markups are homogeneous across producers, equation (24) holds with equality and entry is efficient.[23]

[21] The thought experiments undertaken in this part, by design, cannot inform the equilibrium welfare effects of different policy interventions, they provide key intuitions for the welfare analysis that is to follow.

[22] The conclusion that firms with high markups are inefficiently small is a robust feature of the literature on markups in macroeconomics, and emerges across models of monoplistic competition with variable price elasticities of demand (Dhingra & Morrow, 2019; Baqaee *et al.*, 2022; Edmond *et al.*, 2021), oligopolistic competition, e.g. Atkeson & Burstein (2008), or limit pricing (Peters 2020).

[23] Note that efficiency requires producers of entry and overhead goods to have the same degree of labor market power as final good firms. To see this, consider an economy where markups and markdowns

Selection distortion Consider a marginal increase in the selection cutoff, reallocating the labor freed up by the exiting varieties towards additional entry. If this reallocation increases welfare, selection is too weak.

Lemma 3. *In a given allocation, selection is too weak if, and only if,*

$$\frac{\bar{\epsilon} - \epsilon_{\theta*}}{\bar{\delta}} + \frac{\delta_{\theta*} - \bar{\delta}}{\bar{\epsilon}} \frac{\mathcal{M}_{\theta*}}{\mu_{\theta*}} + \frac{\delta_o - \bar{\delta}}{\bar{\epsilon}} \left(1 - \frac{\mathcal{M}_{\theta*}}{\mu_{\theta*}}\right) > 0, \tag{25}$$

where $\bar{\epsilon} = \mathbb{E}_s\left[\epsilon_\theta\right]$ and $\bar{\delta} = \mathbb{E}_{wn}\left[\delta_\omega\right]$. In the reverse case, selection is too strong. Denoting $\bar{\boldsymbol{\epsilon}} = {}^{\bar{\epsilon}}/_{\bar{\delta}}$ and $\boldsymbol{\mu}^e_\theta \equiv \frac{\mu_\theta}{\mathcal{M}_\theta}$, equation (25) can also be written,

$$\bar{\boldsymbol{\epsilon}} - \boldsymbol{\epsilon}_{\theta*} + (1 - {}^{1}/_{\boldsymbol{\mu}^e_\theta})\bar{\boldsymbol{\epsilon}}\left({}^{\delta_o}/_{\delta_{\theta*}} - 1\right) > 0. \tag{26}$$

Following Lemma 3, selection is too weak if the average entrant is valued more by households than the marginal entrant. The first term in (25) compares consumption gains from raising entry, $\bar{\epsilon} - 1$, to losses from forcing the marginal entrant to exit, $\epsilon_{\theta*} - 1$. The second and third term assess the corresponding change in worker surplus. The homogeneity of effective markups uniquely ensures that selection is efficient and equation (25) holds with equality.

Together, lemmas 1, 2, and 3 characterize distortions in the allocations of a given stock of resources implied by a given allocation. Next, I describe distortions in the supply of resources.

3.3 Factor Supply Distortion

When factors are supplied at an opportunity cost, a lack of competition leads to suboptimal factor provision.When supplied efficiently, the societal costs of factor supply are equated to the marginal rate of transformation between leisure and consumption,

$$-\frac{U_N(\mathcal{Y}^{\text{opt}}, \mathcal{N}^{\text{opt}})}{U_{\mathcal{Y}}(\mathcal{Y}^{\text{opt}}, \mathcal{N}^{\text{opt}})} = \frac{\mathcal{Y}^{\text{opt}}}{\mathcal{N}^{\text{opt}}} \frac{\bar{\epsilon}^{\text{opt}}}{\bar{\delta}^{\text{opt}}}, \tag{27}$$

where $\bar{\delta}^{\text{opt}} \equiv (\int_\omega \Psi'(\frac{n^{\text{opt}}_\omega}{\mathcal{N}^{\text{opt}}})\frac{n^{\text{opt}}_\omega}{\mathcal{N}^{\text{opt}}} dM^{E,\text{opt}}(\omega))^{-1}$ and $\bar{\epsilon}^{\text{opt}} \equiv (\int_\theta \Upsilon'(\frac{y^{\text{mk}}_\theta}{\mathcal{Y}^{\text{opt}}_\theta})\frac{y^{\text{opt}}_\theta}{\mathcal{Y}^{\text{opt}}_\theta} dM^C(\theta))^{-1}$. In contrast, the market allocation solves,

$$-\frac{U_N(\mathcal{Y}^{\text{mkt}}, \mathcal{N}^{\text{mkt}})}{U_{\mathcal{Y}}(\mathcal{Y}^{\text{mkt}}, \mathcal{N}^{\text{mkt}})} = \frac{\mathcal{Y}^{\text{mkt}}}{\mathcal{N}^{\text{mkt}}}. \tag{28}$$

are homogeneous across final good producers, while entry and overhead good producers have no market power, $\beta_o = \beta_e \to \infty$. Equation (24) implies that entry would be excessive in this economy. Intuitively, wages are closer to the opportunity cost of work, which reduces the aggregate magnitude of the non-appropriability externality in labor markets. Thus, the business stealing externality dominates, leading to excessive entry.

Comparing (27) and (28) shows that imperfect competition induces a “labor wedge,” implying suboptimally lower factor supply under standard assumptions.

Proposition 1. *If aggregate labor supply is elastic, the market allocation is inefficient. If* $\forall C, N, 1 + \frac{\partial \log U_{\mathcal{Y}}}{\partial \log N} + \frac{\partial \log U_{\mathcal{Y}}}{\partial \log \mathcal{Y}} > 0$, *labor supply in the planner's equilibrium* N^{opt} *is strictly larger than labor supply in the market equilibrium* $\mathcal{Y}^{mkt}$.

Distortions in factor supply are conceptually distinct from those highlighted in lemmas 1, 2, and 3. That is, the misalignment of private and social returns to supplying factors leads to a deadweight loss in the decentralized economy, irrespective of the efficiency at which factors are allocated across productive uses.

While conceptually distinct, the positive and normative implications of factor supply and allocative distortions are deeply intertwined. Intuitively, a reduction in the supply of factors, i.e., the stock of productive resources, due to fixed costs, induces higher barriers to firm entry and lower competition in the economy. Falling competitive pressures lead to resource reallocations the micro-level, and changes in allocative efficiency, total factor productivity, and, ultimately, the stock of resources at the macro-level. The next section unbundles the implications of micro-level interactions between firms' product and factor market power for the macro-level response of total factor productivity and welfare to changes in market size.

4 Unbundling Market Power: Market Size and Aggregate Productivity

In this section, I delineate the importance of unbundling firms' effective market power for describing the equilibrium response of aggregate outcomes to policy intervention or, more broadly, changes in the economic environment.

4.1 Firm-level sufficient statistics

I begin by characterizing how firm-level prices and quantities respond to equilibrium shifts in market aggregates and exogenous shocks.

Lemma 4. *Equilibrium changes in the relative price* $\hat{p}_\theta = \frac{p_\theta}{\mathcal{P}}$ *and quantitiy* $\hat{y}_\theta = \frac{y_\theta}{\mathcal{Y}}$ *of a firm of type* θ *can be written,*

$$d \ln \hat{p}_\theta = -\Gamma_\theta d \ln A_\theta + \underbrace{(1 - \boldsymbol{\Gamma}_\theta) d \ln \boldsymbol{\mathcal{P}}}_{\Delta \boldsymbol{\mu}^e_\theta} - \underbrace{\frac{\gamma_\theta - \boldsymbol{\Gamma}_\theta}{\beta_\theta \gamma_\theta} d \ln \mathcal{A}}_{\Delta w_\theta}, \tag{29}$$

$$d\ln\hat{y}_\theta = \boldsymbol{\Sigma}_\theta\boldsymbol{\Gamma}_\theta\left(d\ln A_\theta + d\ln\boldsymbol{\mathcal{P}}\right) - \frac{\boldsymbol{\Sigma}_\theta}{\beta_\theta}\frac{\boldsymbol{\Gamma}_\theta}{\gamma_\theta}d\ln\boldsymbol{\mathcal{A}}, \tag{30}$$

where $d\ln\boldsymbol{\mathcal{P}} \equiv -d\ln\frac{\mathcal{W}}{\mathcal{P}}$ *and* $d\ln\mathcal{A} = d\ln\frac{C}{N}$*,* $\boldsymbol{\Sigma}_\theta \equiv \rho_\theta\gamma_\theta\frac{\beta_\theta+\sigma_\theta}{\gamma_\theta\beta_\theta+\rho_\theta\sigma_\theta}$ *denotes a firm's effective cost price pass-through and* $\boldsymbol{\Sigma}_\theta = \frac{\beta_\theta\sigma_\theta}{\beta_\theta+\sigma_\theta}$ *its effective price elasticity of demand..*

Lemma 4 shows that changes in firm-level prices and quantities can be, respectively, decomposed into three terms. As I will now explain, two terms capture the extend to which firm behavior is isomporhic to that predicted by workhorse models of variable markups (markdowns) with competitive factor (product) markets. Hence, new ramifications that arise unqiuely from firm-level interactions between markups and markdowns are captured by the final term.

The first two terms on the right-hand-side of (29) and (30) show that a firm's response to idiosyncratic shocks $d\ln A_\theta$ and economy-wide competition $d\ln\boldsymbol{P}$ can be summarized by a single elasticity $\boldsymbol{\Gamma}_\theta$, which I refer to as a firm's effective cost pass-through. Similarly, a firms' effective price elasticity of demand $\boldsymbol{\Sigma}_\theta$ captures how its output changes in response to idiosyncratic cost and competition shocks. As a key implication, this shows that canonical models of variable markups are nested by the model, and that existing estimates of price cost pass-throughs, in principle, remain informative of key structural elasticities.[24]

However, Lemma 4 shows that firms endowed with both price and wage-setting power respond not only to competitive pressures, but also to shifts in aggregate factor productivity, $d\ln\mathcal{A}$. Intuitively, both factor and product market power can effectively insulate a firm from competitive pressures in the product market. However, high factor market power effectively lowers a firm's scale inefficiency. Formally, the importance of cost relative to competition shocks for price changes scales with the difference between a firm's structural wage pass-through and its effective pass-through $\gamma_\theta - \Gamma_\theta$, and for quantities with the ratio of its effective demand and structural factor supply elasticity.

Effective pass-throughs and demand elasticities turn out to be summary statistics for describing the behavior of firms that excert either price or wage-setting power. For the purpose of understanding how interactions between market powers matter for aggregate outcomes in one economy, it is useful to define economies that feature no such interactions but are "obervationally equivalent" in terms of these sufficient statistics.

Definition 1 *If* $\mathcal{X}$ *is an equilibrium allocation and* $\mathcal{T} = (\{\Psi_\omega\}_\Omega, \{\Upsilon_\theta\}_\Theta, f_e, f_o)$ *,let* $\Xi(\mathcal{X},\mathcal{T}) = \left(\left\{\boldsymbol{\mu}^e_\theta(\frac{n_\theta}{\mathcal{N}}), \boldsymbol{\Gamma}_\theta(\frac{n_\theta}{\mathcal{N}}), \boldsymbol{\Sigma}_\theta(\frac{n_\theta}{\mathcal{N}})_\Theta\right\}, \bar{\boldsymbol{\epsilon}}\right)$*. An economy* $\mathcal{T}'$ *is markup-equivalent to* $\mathcal{T}$*, if* $\Xi(\mathcal{X},\mathcal{T}') = (\{\mu_\theta, \rho_\theta, \sigma_\theta\}_\Theta, \bar{\boldsymbol{\epsilon}}) = \Xi(\mathcal{X},\mathcal{T})$*.*

[24] For intuition, note that as $\beta_\theta \to \infty, \gamma_\theta \to 1$, a firm's price response takes the form $d\ln\hat{p}_\theta = (1-\rho_\theta)d\ln\mathcal{P} - \rho_\theta d\ln A_\theta$, which, up to first order, is implied by a wide class of demand systems and models of imperfect competition. See Amiti *et al.* (2019) for a moe detailed discussion.

4.2 Aggregate Productivity and Effective Market Size

The response of aggregate factor productivity, $\mathcal{A} = \frac{\mathcal{Y}}{\mathcal{N}}$, to a change in market size is central for welfare and policy analysis. Regarding welfare, it informs the costs of factor supply distortions; moreover, it also informs the the policy returns to entry subsidies, trade integration, or population growth.

Remark. If $d\ln f_e = d\ln f_o = d\ln f$, then $\frac{\partial \ln \boldsymbol{\mathcal{A}}}{\partial \ln \mathcal{N}} = \frac{\partial \ln \boldsymbol{\mathcal{A}}}{\partial \ln L} = -\frac{\partial \ln \boldsymbol{\mathcal{A}}}{\partial \ln f}$.

The following characterizes the response of aggregate productivity to changes in the economy's effective market size, $\frac{d\ln \mathcal{A}}{d\ln \mathcal{L}} = d\ln \mathcal{N}Lf$, where $d\ln f$ is a proportional change in entry and overhead requirements. If factor supply is inelastic, this captures the response of welfare to an exogenous shock to market size. If factor supply is elastic, this elasticity informs the equilibrium comovement of factor supply and productivity.

Theorem 2. *In response to an increase in effective factor supply $d\ln \mathcal{L} = d\ln \mathcal{N}Lf$, the change in total factor productivity $d\ln \boldsymbol{\mathcal{A}} = d\ln \frac{\mathcal{Y}}{\mathcal{N}}$ is given by,*

$$\frac{d\ln \mathcal{A}}{d\ln \mathcal{L}} = (\bar{\boldsymbol{\epsilon}} - 1) + \frac{\left(\boldsymbol{\zeta}^{\approx} + \zeta_{\theta^*}\right)\left(\bar{\boldsymbol{\epsilon}} - \frac{1}{\beta_F}\right) + \left(\zeta_{\boldsymbol{\Sigma}} + \zeta_{\boldsymbol{\Gamma}} + \zeta_{\mathcal{E}}\right)\left(\bar{\boldsymbol{\epsilon}}\right)}{1 - \boldsymbol{\zeta}^{\approx} - \zeta_{\boldsymbol{\Sigma}} - \zeta_{\boldsymbol{\Gamma}} - \zeta_{\theta^*} - \zeta_{\mathcal{E}}}, \tag{31}$$

where

$$\begin{aligned}
\boldsymbol{\zeta}^{\approx} &= (\bar{\boldsymbol{\epsilon}} - 1)Cov_s\left[\boldsymbol{\Sigma}_\theta, {}^{1}/_{\boldsymbol{\mu}^e_\theta}\right] + \mathbb{E}_s\left[(1 - {}^{\bar{\epsilon}}/_{\boldsymbol{\mu}^e_\theta})(1 - \boldsymbol{\Gamma}_\theta)\boldsymbol{\Sigma}_\theta\right]\boldsymbol{\Lambda} + s_{\theta^*}\iota_{\theta^*}\left(\bar{\boldsymbol{\epsilon}} - \bar{\boldsymbol{\epsilon}}_{\theta^*}\right)\frac{\boldsymbol{\Lambda} - \Lambda_{\theta^*}}{\Lambda_{\theta^*}} \\
\zeta_{\boldsymbol{\Sigma}} &= (\bar{\boldsymbol{\epsilon}} - 1)\left\{Cov_s\left[{}^{\Sigma_\theta}/_{\beta_\theta} - \mathcal{M}_\theta, {}^{1}/_{\mu_\theta}\right] + Cov_s\left[{}^{1}/_{\boldsymbol{\mu}^e_\theta}, {}^{\Sigma_\theta}/_{\sigma_\theta\beta_\theta}\right]\right\} \\
\zeta_{\boldsymbol{\Gamma}} &= \mathbb{E}_s\left[(1 - {}^{\bar{\epsilon}}/_{\boldsymbol{\mu}^e_\theta}){}^{(\gamma_\theta - \boldsymbol{\Gamma}_\theta)}/_{\beta_\theta\gamma_\theta}\right]\Lambda_\mu + \mathbb{E}_s\left[(1 - {}^{\bar{\epsilon}}/_{\boldsymbol{\mu}^e_\theta})\boldsymbol{\Sigma}_\theta(1 - \boldsymbol{\Gamma}_\theta)\right](\Lambda_\mu - \boldsymbol{\Lambda}) \\
\zeta_{\theta^*} &= s_{\theta^*}\iota_{\theta^*}\left({}^{\delta_o}/_{\delta_{\theta^*}} - 1\right)\left(\boldsymbol{\Lambda} - \Lambda_{\theta^*}\right) \\
\zeta_{\mathcal{E}} &= \mathbb{E}_s\left[(1 - {}^{\bar{\epsilon}}/_{\boldsymbol{\mu}^e_\theta})\right]\mathbb{E}_s\left[{}^{\mathcal{M}_\theta}/_{\bar{\mathcal{M}}_F} - {}^{\mu_\theta - 1}/_{\mu_\theta - \mathcal{M}_\theta}\right]
\end{aligned}$$

and $\boldsymbol{\Lambda} \equiv \mathbb{E}_s\left[1 - \frac{\mathcal{M}_\theta}{\mu_\theta}\right]$, $\Lambda_\mu = \mathbb{E}_s\left[1 - \frac{1}{\mu_\theta}\right]$, ${}^{1}/_{\beta_F} = \mathbb{E}_{wn}\left[\frac{1}{\beta_\omega}|\omega \in \{o, e\}\right]$, $\bar{\boldsymbol{\epsilon}} = \frac{\bar{\epsilon}}{\delta}$ *and* ${}^{1}/_{\iota_{\theta^*}} = \left[\frac{1 - G(\theta)}{g(\theta)}\frac{\partial \ln X_\theta}{\partial \theta}\right]_{\theta^*} > 0.$

Theorem 2 parses how firm heterogeneity and interactions between factor and product markets jointly determine the response of aggregate productivity to a change in market size. Specifically, the first term on the right-hand side of (31) captures the response of a counterfactual economy with the same effective markups and efficient allocations. Consequently, the second term captures and decomposes the aggregate returns to market size arising from changes in allocative efficiency. In what follows, I briefly describe the first, and then discuss the second term in greater detail.

Capturing what Baqaee & Farhi (2020) refer to as a technical efficiency effect, the first term in (31) captures aggregate returns to scale from fixed costs. Holding fixed allocations, an increase in effective market size, $d\ln \mathcal{L}$, first raises entry, $d\ln M = \frac{1}{\delta}d\ln \mathcal{L}$, then output, proportionally to $d\ln\mathcal{Y} = \bar{\boldsymbol{\epsilon}}d\ln L\mathcal{N}$, and, finally, factor productivity by $d\ln\boldsymbol{A} = (\bar{\boldsymbol{\epsilon}} - 1)d\ln\mathcal{L}$. Conditional on households' effective surplus $\bar{\epsilon}$, the origins of firms' market power are irrelevant for describing this effect. Consequently, theorems 1 and 2 jointly imply that distinguishing markups from markdowns is immaterial for welfare and policy analysis when resource allocations are efficient.

The remaining terms in (31) describe how resource reallocations between firms contribute to aggregate productivity, and distentangle the implications of micro-level interactions between markups and markdowns. Specifically, by setting ${}^{1}/_{\beta_F=\zeta_{\theta^*}} = \zeta_{\boldsymbol{\Sigma}} = \zeta_{\boldsymbol{\Gamma}} = \zeta_{\mathcal{E}} = 0$, Theorem 2 describes the productivity response of a markup-equivalent economy, following definition 1. The term $\zeta_{\theta^*}$and scalar ${}^{1}/_{\beta_F}$ describe changes in allocative efficiency that uniquely occur in a markdown-eqivalent economy. Finally, the terms ζ_{μ}, $\zeta_{\mathcal{E}}$ and $\zeta_{\boldsymbol{\Gamma}}$ capture changes in allocative efficiency that are due to firm-level interactions between product and factor market power.

In the following, I provide a number of illustrative examples to further highlight the economic insights implied by each of the terms in Theorem 2.

Markup vs markdown-equivalent economies The term $\boldsymbol{\zeta}^{\approx}$ describes the response of consumer's surplus, holding fixed factor prices.

Corollary 1. *In a markup-equivalent economy, the elasticity of aggregate productivity to changes in effective market size* $\mathrm{d}\ln\mathcal{L} = d\ln\mathcal{N}Lf$ *is given by,*

$$\frac{d\ln\boldsymbol{\mathcal{A}}}{d\ln\mathcal{L}} = (\bar{\boldsymbol{\epsilon}} - 1) + \frac{\boldsymbol{\zeta}^{\approx}}{1-\boldsymbol{\zeta}^{\approx}}\bar{\boldsymbol{\epsilon}}, \tag{32}$$

The expression in (32) is isomporphic to the gains from population growth in an economy with competitive factor markets and inelastic labor supply, analyzed by Baqaee *et al.* (2022). Following Theorem 2, the economic forces that govern $\boldsymbol{\zeta}^{\approx}$ may stem from either factor and product market power. It is hence intuitive that (32), at least partially, captures the aggregate returns from market size in a markdown-equivalent economy, as shown by the following.

Corollary 2. *In a markdown-equivalent economy where* $\mathbb{E}_s\left[\mathcal{M}_\theta\right] = \frac{\beta_F}{\beta_F+1}$*, the elasticity of aggregate productivity to changes in effective market size* $\mathrm{d}\ln\mathcal{L} = d\ln\mathcal{N}Lf$ *is given by,*

$$\frac{d\ln\boldsymbol{\mathcal{A}}}{d\ln\mathcal{L}} = (\bar{\boldsymbol{\epsilon}} - 1) + \frac{\boldsymbol{\zeta}^{\approx} + \zeta_{\theta^*}}{1-\boldsymbol{\zeta}^{\approx} - \zeta_{\theta^*}}\left(\bar{\boldsymbol{\epsilon}} - {}^{1}/_{\beta_F}\right), \tag{33}$$

Conditional on equilibrium aggregates, variety producers in markup- and markdown-equivalent economies behave symmetrically. It is, thus, intuitive that differences the aggregate returns to market size implied by (33) and (32) stem from extensive margin adjustments. Mechanically, the costs of generating additional entrants are higher in a markup-equivalent economy, and hence the change in allocative efficiency scales with $(\bar{\boldsymbol{\epsilon}} - \frac{1}{\beta_f})$ compared to $(\bar{\boldsymbol{\epsilon}})$ in a markup-equivalent economy.

Less mechanically, additional effects arises from the selection margin, captured by ζ_{θ^*}. Following Lemma 3, tougher selection raises allocative efficiency if the average entrant generates a higher social surplus than the marginal entrant. While the term $\boldsymbol{\zeta}^{\approx}$ accounts for differences in effective consumption surpluses via $\bar{\boldsymbol{\epsilon}} - \boldsymbol{\epsilon}_{\theta^*}$, in a markdown-equivalent economy such comparisons require accounting for the relative surplus generated by overhead sector jobs, captured by the term ${}^{\delta_o}/_{\delta_{\theta^*}} - 1$.

Heterogeneity in markups and markdowns When power stems from both factor and product markets, Lemma 4 implies that effective markups and pass-throughs are insufficient to characterize the equilibrium behavior of firms. The term $\zeta_{\boldsymbol{\Sigma}}$ captures the implied changes in allocative efficiency attributable to heterogeneity in the relative markdowns and markups across firms. Following Baqaee *et al.* (2022), the following example illustrates $\zeta_{\boldsymbol{\Sigma}}$.

Corollary 3. *Suppose that for all* θ, $\Upsilon_\theta(\frac{y_\theta}{\mathcal{Y}}) = (\frac{y_\theta}{\mathcal{Y}})^{\frac{\sigma_\theta - 1}{\sigma_\theta}}$, $\Psi_\theta(\frac{n_\theta}{\mathcal{N}}) = (\frac{n_\theta}{\mathcal{N}})^{\frac{\beta_\theta + 1}{\beta_\theta}}$, $f_o = 0$, *and* $\Psi_e(\frac{n_\theta}{\mathcal{N}}) = (\frac{n_\theta}{\mathcal{N}})^{\frac{\beta_e + 1}{\beta_e}}$. *Then, the elasticity of aggregate productivity to changes in effective market size* $d\ln \mathcal{L} = d\ln \mathcal{N} L f$ *is given by,*

$$\frac{d\ln \boldsymbol{\mathcal{A}}}{d\ln \mathcal{L}} = (\bar{\boldsymbol{\epsilon}} - 1) + \frac{\boldsymbol{\zeta}^{\approx}\left(\bar{\boldsymbol{\epsilon}} - \mathbb{E}_s\left[{}^{1}/_{\beta_\theta}\right]\right) + \left(\zeta_{\boldsymbol{\Sigma}} + \zeta_{\mathcal{E}}\right)\left(\bar{\boldsymbol{\epsilon}}\right)}{1 - \boldsymbol{\zeta}^{\approx} - \zeta_{\boldsymbol{\Sigma}} - \zeta_{\mathcal{E}}},$$

where $\boldsymbol{\zeta}^{\approx} = (\bar{\boldsymbol{\epsilon}} - 1) Cov_s\left[\boldsymbol{\Sigma}_\theta, {}^{1}/_{\boldsymbol{\mu}^e_\theta}\right]$.

Generalized CES preferences imply heterogeneous, but exogenous markups and markdowns, hence $\zeta_{\boldsymbol{\Gamma}} = 0$. In the absence of overhead requirements, $f_o = 0$, selection does not contribute to allocative efficiency, $\zeta_{\theta^*} = 0$. Following (30), firms with lower effective demand elasticities $\boldsymbol{\Sigma}_\theta$ are relatively more shielded from competition. It is easy to check that such firms also have higher effective markups.[25] Following Lemma 1, competition hence raises allocative efficiency by reallocating factors towards firms that were initially more distorted, captured by $\boldsymbol{\zeta}^{\approx} = Cov_s\left[\boldsymbol{\Sigma}_\theta, {}^{1}/_{\boldsymbol{\mu}^e_\theta}\right] > 0$.[26]

As argued before, rising aggregate productivity separately affects firms' quantity decisions. Whether this contributes or works against the gains implied by competition hinges on how markups and markdowns are distributed within firms. If markdowns are the primary driver of a firm's low effective demand elasticity, the gains implied by competition ought to be corrected downward, and vice versa. Thereby, this term

[25] Note that $\boldsymbol{\Sigma}_\theta = \frac{\mu_\theta \mathcal{M}_\theta}{\mu_\theta - \mathcal{M}_\theta}$.

[26] Baqaee *et al.* (2022) call this the Darwinian effect.

captures the simple intuition that firms that exploit factors to create consumer surplus are less valuable to society, highlighting why unbundling markups is important.

In summary, while competition always induces efficiency-improving reallocations from firms with low to those with high effective markups when one market is competitive, interactions between markups and markdowns imply that such gains can no longer be guaranteed, but may also be larger.

The importance of unbundling pass-throughs. The term $\zeta_{\boldsymbol{\Gamma}}$captures how interactions between markups and markdowns shape the pro-competitive effects of increasing the size of the market. The intuitions are similar to those put forward to argue why unbundling markups is important. Following Lemma 4, if higher cost pass-throughs Γ_θ are associated with higher wage pass-throughs γ_θ, reductions in effective markups following an increase in competitive pressures in the product market may not translate into lower consumer prices if wages respond even more to factor market competition. Whether or not this raises allocative efficiency, in turn, depends on the intial efficiency of entry, as well as the relative gains in the surplus households derive in the factor market.

5 Welfare

Utilizing the insights from the previous section, I now characterize the economy's response to efficiency-inducing policies, which coincides with the social costs of distortions. To that end, denote, again, the allocation vector $\mathcal{X} = (\frac{n_e}{\mathcal{N}}, \frac{n_e}{\mathcal{N}}, \{\frac{n_\theta}{\mathcal{N}}\}_{\theta\in\Theta})$ to describe how a given supply of labor $\mathcal{N}$ is allocated across entry, overhead and final good production. Given an allocation $\mathcal{X}$ that is feasible for $\mathcal{N}$, let $\mathcal{U}(N, \mathcal{X})$ denote the implied level of household welfare. Then, denote $d\tau^{\mathcal{X}} = d\mathcal{X} \equiv \mathcal{X}^{\text{opt}}/\mathcal{X}^{\text{mkt}}$ the vector of deviations of allocations from their value at the efficient allocation, and $d\ln\tau^N \equiv \ln\mathcal{N}^{\text{opt}}/\mathcal{N}^{\text{mkt}}$ the log-deviation of factor supply from its efficient value. The following characterizes the distance to the efficient frontier, up to second order.

Proposition 2. *The distance to the efficient frontier, in welfare equivalent terms and up to second order, can be approximated as,*

$$\mathcal{L} \approx \underbrace{\frac{1}{2}\eta\frac{\partial\ln\mathcal{U}}{\partial\ln\mathcal{N}}d\ln\tau^N}_{\textit{Deadweight Loss}} + \overbrace{\underbrace{\frac{1}{2}\eta\frac{\partial\ln\mathcal{U}}{\partial\mathcal{X}}\frac{\partial\mathcal{X}}{d\ln\tau^N}d\ln\tau^N}_{\textit{Indirect effect}} + \underbrace{\frac{1}{2}\eta\frac{\partial\ln\mathcal{U}}{\partial\mathcal{X}}d\tau^{\mathcal{X}}}_{\textit{Direct effect}}}^{\textit{Allocative Efficiency Loss}} \tag{34}$$

where $\eta \equiv (\frac{\partial\ln U(C^{mkt},N^{mkt})}{\partial\ln C})^{-1}$, $\bar{\boldsymbol{\epsilon}} \equiv \frac{\mathbb{E}_s[\epsilon_\theta]}{\mathbb{E}_{wn}[\delta_\omega]}$, Proposition 2 shows that welfare losses occur alongside three margins. The first two terms in (34) capture the forces described in Theorem 2 - that is the direct and indirect contribution of factor supply distortions to welfare losses. The third term captures allocative efficiency losses, holding fixed the economy's stock of resources at the initial market allocation. Thereby, Proposition 2 provides important insights about the limitations of targeted firm-level policies as a tool to counter welfare losses of imperfect competition. Up to second order, firm-specific taxes that solely seek to correct allocative inefficiencies leave distortions in factor supply unchanged. Thus, distinguishing between direct and indirect allocative efficiency losses is helpful in informing the effectiveness of targeted policy interventions. Given that I now turn to characterizing these effects analytically.

5.1 Direct Allocative Efficiency Loss

To characterize the direct welfare loss from factor misallocations, I study the change in consumer surplus in response to optimal firm-level policies. Appendix A.4 provides an example of a tax scheme that achieves this. To remove distortions in relative firm sizes, the policy sets markups equal to the consumer surplus and markdowns equal to the worker surplus generated by each entrant, $\mu_\theta^{opt} = \epsilon_\theta$ and $\mathcal{M}^{\text{opt}}{}_\theta = \delta_\theta$, while incentivizing entry through sales and wage bill subsidies equalling $\tau_\theta^s = 1/\mu_\theta$ and $\tau_\theta^{wn} = 1/\mathcal{M}_\theta$. To induce optimal incentives on the extensive margin, the prices of overhead and entry inputs are taxed a rate $\tau_e = \overline{\delta}/\delta_e$ and $\tau_o = \overline{\delta}/\delta_o$. Lump-sum taxes on households balance the budget.

The following characterizes the welfare change induced by such a policy.

Proposition 3. *The consumption loss implied by the direct efficiency effect is given by,*

$$\frac{\partial \ln \mathcal{Y}}{\partial \mathcal{X}} d\tau^{\mathcal{X}} = \mathbb{E}_s \left[\boldsymbol{\Sigma}_\theta \left({}^{\bar{\epsilon}}/_{\boldsymbol{\mu}_\theta^e} - \bar{\boldsymbol{\epsilon}} \mathbb{E}_s \left[{}^{1}/_{\boldsymbol{\mu}_\theta^e} \right] \right)^2 \right] + \mathbb{E}_s \left[\boldsymbol{\Sigma}_\theta \right] \left(\mathbb{E}_s \left[{}^{\bar{\epsilon}}/_{\boldsymbol{\mu}_\theta^e} \right] - 1 \right)^2 \\ + s_{\theta^*} \iota_{\theta^*} \left(\overline{\boldsymbol{\epsilon}} - \boldsymbol{\epsilon}_{\theta^*} + (1 - {}^{1}/_{\boldsymbol{\mu}_\theta^e}) \overline{\boldsymbol{\epsilon}} \left({}^{\delta_o}/_{\delta_{\theta^*}} - 1 \right) \right)^2$$

Proposition 3 shows how each each source of allocative inefficiency highlighted in section 3 contributes to welfare losses: Variable labor allocations, entry, and selection.

The first term captures distortions in variable factor allocations (Lemma 1). On the one hand, it scales with the dispersion in the ratios of firm rents relative to the average rents earned by households ${}^{\bar{\epsilon}}/_{\mu_\theta^e}$. It also scales with the elasticities of product demand and labor supply. Efficiency losses scale also the effective price elasticity of demand $\boldsymbol{\Sigma}_\theta$ as removing a given amount of dispersion in effective markups $\boldsymbol{\mu}_\theta^e = \mu_\theta/\mathcal{M}_\theta$ requires higher subsidies to sales and wage bills and, thus, larger offsetting lump-sum taxes on households if demand and factor supply, generally, respond more strongly to changes in prices.

The second term captures distortions from inefficient entry (see Lemma 2). The costs of such distortions are higher the larger the distance between the social and private returns to entry, $\left|\mathbb{E}_s\left[\bar{\epsilon}/\mu^e_\theta\right]-1\right|$, and the effective price elasticity of demand Σ_θ captures the relevant elasticity of the entry margin if the supply of factors is held fixed.

The last term quantifies the societal costs of inefficient selection (Lemma 3). It scales with the squared difference between the infra-marginal household surpluses of the marginal and average firm. The costs of selection inefficiencies are increasing in the sales share of marginal firms, and the sensitivity of exit behavior to distortions (ι_{θ^*}).

Note that Proposition 3 implies that unbundling firms' effective markups is unnecessary to describe the second-order welfare implications of budget neutral firm-level tax interventions. Intuitively, following the discussion in the previous section, sufficient statistics capturing changes in consumer surplus in response to competition and firm-level shocks are sufficient to describe the implications of shocks that leave factor market competition unchanged.

5.2 Total Welfare Loss

The costs of factor supply distortions are closely tied to the response of aggregate productivity to market size given by Proposition 3. Since factor supply is given by the implicit function $-\frac{U_N(\mathcal{Y},\mathcal{N})}{U_C(\mathcal{Y},\mathcal{N})}-\frac{\mathcal{Y}}{\mathcal{N}}(1+\tau)=0$, changes in welfare from a change in subsidies to households' earnings $d(1+\tau)$, are given by

$$d\ln\mathcal{U}=\frac{\partial\ln U}{\partial\ln\mathcal{Y}}\left(((1-(1+\tau)\frac{\partial\ln\mathcal{N}}{\partial\ln\mathcal{A}})\frac{d\ln\mathcal{A}}{d\ln\mathcal{N}}+1)\frac{d\ln\mathcal{N}}{d(1+\tau)}-(1+\tau)\frac{\partial\ln\mathcal{N}}{\partial(1+\tau)}\right)d(1+\tau),$$

where $\frac{\partial\ln\mathcal{N}}{\partial(1+\tau)}=\frac{1}{\frac{\partial\ln U_1U_2}{\partial\ln\mathcal{Y}}-\frac{\partial\ln U_1U_2}{\partial\ln\mathcal{N}}}$, $\frac{\partial\ln\mathcal{N}}{\partial\ln\mathcal{A}}=\frac{\partial\ln U_1U_2}{\partial\ln\mathcal{Y}}\cdot\frac{\partial\ln\mathcal{N}}{\partial(1+\tau)}$ and $\frac{d\ln\mathcal{N}}{d(1+\tau)}=\frac{1}{1-\frac{\partial\ln\mathcal{N}}{\partial\ln\mathcal{A}}\frac{d\ln\mathcal{A}}{d\ln\mathcal{N}}}\cdot\frac{\partial\ln\mathcal{N}}{\partial(1+\tau)}$ all depend only on aggregates and the elasticities of $U(.,.)$. The following characterizes the distance to the efficient frontier using Theorem 2 for GHH preferences.

Proposition 4. *If $U(\mathcal{Y},\mathcal{N})=\log\left(\mathcal{Y}-\psi\frac{N^{1+1/\varphi}}{1+1/\varphi}\right)$, the distance to the efficient frontier in consumption equivalents, up to second order, can be approximated by*

$$\mathcal{L}^{CE}\approx\frac{1}{2}\cdot\frac{\varphi\varepsilon^{\mathcal{A}}_{\mathcal{L}}}{1-\varphi\varepsilon^{\mathcal{A}}_{\mathcal{L}}}\left(\bar{\boldsymbol{\epsilon}}-1\right)+\frac{1}{2}\frac{\partial\ln\mathcal{Y}}{\partial\mathcal{X}}d\tau^{\mathcal{X}}. \qquad (35)$$

where $\varepsilon^{\mathcal{A}}_{L}=\frac{d\ln\mathcal{A}}{d\ln\mathcal{L}}$ is given by Theorem 2 and $\frac{\partial\ln\mathcal{Y}}{\partial\tau^{\mathcal{X}}}d\tau^{\mathcal{X}}$ by Proposition 2.

6 Quantification

This section applies the theoretical results to quantify and decompose the societal costs of imperfect competition. I begin by outlining and implementing a non-paramaetric approach to calibrating the firm-level statistics highlighted by the theory.

6.1 Calibration Approach

I build on Baqaee *et al.* (2022) to calbrate the model.[27] My setting requires adjusting their approach, which requires joint estimates of price and wage pass-throughs, which, to the best of my knowledge, do not exist. More substantially, commonly deployed strategies estimate wage and price pass-throughs from changes in either prices or wages in response to plausibly exogenous firm-level shocks. Such an approach does not separately identify the structural elasticities ρ_θ and γ_θ when both markups and markdowns are endogenous. Appendix C further elaborates on this point.

To overcome these challenges, I show how relatively mild restrictions on the functional form of firms' labor supply and product demand in combination with separate cross-sectional estimates of markups and markdowns can be used to infer the structural pass-throughs ρ_θ and γ_θ, as well as consumer and worker rents ϵ_θ and δ_θ.

Functional form restrictions For the baseline caloibration, I assume GHH household preferences over consumption and leisure. For robustness, I also report results for KPR preferences that imply no wealth effects on labor supply.

$$U_{GHH}(\mathcal{Y}, N) = \log\left(\mathcal{Y} - \psi \frac{\mathcal{N}^{1+1/\varphi}}{1+1/\varphi}\right), \quad U_{KPR}(\mathcal{Y}, N) = \log \mathcal{Y} - \psi \frac{\mathcal{N}^{1+1/\varphi}}{1+1/\varphi} \tag{36}$$

In the baseline calibration, I assume that the Frisch elasticity equals $\varphi = 0.5$ following Keane & Rogerson (2012). I normalize the labor disutility parameter ψ, together with the entry cost f_e to achieve output $C = 1$ and a total mass of entrants $M = 1$ at the initial equilibrium.

Residual demand and factor supply curves faced by firms are

Assumption 1 *Firm types θ lie in the unit interval $[0,1]$. For all $\theta \in \Theta$, and $\omega \in \Theta \cup \{o, e\}$, preferences are given by,*

$$\Psi_\omega(\frac{n}{N}) = \Psi(\frac{1}{A^n_\omega}\frac{n_\theta}{N}), \qquad \Upsilon_\theta(\frac{y}{\mathcal{Y}}) = \Upsilon\left(A^c_\theta \frac{y}{\mathcal{Y}}\right), \tag{37}$$

[27] Baqaee *et al.* (2022) leverage estimates of price cost pass-throughs ρ_θ across the sales distribution of Belgian firms by Amiti *et al.* (2019) to infer markups and consumer rents under mild parametric restrictions on the firm-level demand system.

and imply incomplete pass-throughs, that is, $\Psi \in \{\tilde{\Psi} \in C^3 : \Psi' > 0, \Psi'' > 0, \forall x \in \mathbb{R}_0^+, \gamma(x) \in (0,1]\}$ *and* $\Upsilon \in \{\tilde{\Upsilon} \in C^3 : \Upsilon' > 0, \Upsilon'' < 0, \forall x \in \mathbb{R}_0^+, \rho(x) \in (0,1]\}$*.* A_θ^n *and* A_θ^c *denote labor supply and product demand shifters.*

The assumption that the support of the type distribution G equals the unit interval is without loss of generality. The restriction on preferences implies that firms face similarly shaped factor supply and product demand curves, conditional on the realizations of A_θ^c and A_θ^n. These shifters accommodate the possibility that wages, prices, sales, and wage bills are imperfectly correlated across firms, as commonly observed in the data. Differences in product quality, workplace amenities and productivities are not identified separately. [28] However, their product $\log \tilde{A}_\theta = \log A_\theta A_\theta^n A_\theta^c$ can be identified. Since the above functional form restrictions, profits, sales, as well as wage bills are strictly increasing in $\tilde{A}_\theta$,[29] I henceforth refer to $\log \tilde{A}_\theta$ as "productivity." As a result, a firm's type θ can be identified by its position in either the sales or wage distribution.

Pass-throughs In the cross-section of final good producers, the model implies the following relationship between sales shares s_θ, wage bill shares ω_θ^{wn} and firm type θ :

$$\frac{d \log s_\theta}{d\theta} = \frac{\chi_\theta}{\mu_\theta} \frac{d \log \tilde{A}_\theta}{d\theta}, \qquad \frac{d \log \omega_\theta^{wn}}{d\theta} = \frac{\mu_\theta}{\mathcal{M}_\theta} \frac{d \log s_\theta}{d\theta},$$

where $\chi_\theta \equiv \frac{\sigma_\theta \beta_\theta \gamma_\theta \rho_\theta}{\sigma_\theta \rho_\theta + \beta_\theta \gamma_\theta}$. Hence, markdowns and markups comove with sales in the cross-section as follows,

$$\frac{d \log \mu_\theta}{d\theta} = \chi_\theta \frac{(1-\rho_\theta)}{\rho_\theta \sigma_\theta} \frac{d \log \tilde{A}_\theta}{d\theta} = (\mu_\theta - 1) \frac{1-\rho_\theta}{\rho_\theta} \frac{d \log s_\theta}{d\theta}, \tag{38}$$

$$\frac{d \log \mathcal{M}_\theta}{d\theta} = \frac{\mu_\theta}{\mathcal{M}_\theta} (1 - \mathcal{M}_\theta) \frac{\gamma_\theta - 1}{\gamma_\theta} \frac{d \log s_\theta}{d\theta}. \tag{39}$$

Given information on sales shares $\{s_\theta\}_\theta$, as well as markdowns and markups $\{\mathcal{M}_\theta, \mu_\theta\}_\theta$, one can use the differential equations (38) and (39) to recover wage pass-throughs γ_θ and price pass-throughs ρ_θ. This also allows to recover density G of $\tilde{A}_\theta$ up to the normalization that $\tilde{A}_0 = 1$.

Household rents and selection cutoff To infer consumer rents, I use the following model-implied relationship between infra-marginal consumption surpluses ϵ_θ, sales

[28] That is, the model is isomporhic to one with only productivity differences. To see this, define $\tilde{n}_\theta = \frac{n_\theta}{A_\theta^n} = \frac{c_\theta}{A_\theta^n A_\theta} = \frac{1}{A_\theta^n A_\theta A_\theta^c} C \left(\Upsilon'\right)^{-1} \left(\frac{p_\theta}{\mathcal{P}}\right), \tilde{c}_\theta = A_\theta^c c_\theta = A_\theta^c A_\theta n_\theta = A^c A_\theta A_\theta^n N \left(\Psi'\right)^{-1} \left(\frac{w_\theta}{\mathcal{W}}\right). \tilde{y}_\theta = A_\theta n_\theta = A^C A_\theta A_\theta^n \frac{C}{N} \frac{\left(\Upsilon'\right)^{-1}\left(\frac{p_\theta}{\mathcal{P}}\right)}{\frac{N}{C}\left(\Psi'\right)^{-1}\left(\frac{w_\theta}{\mathcal{W}}\right)}$. Redefining $\tilde{A}_\theta = A^c A_\theta A_\theta^n$,

[29] Since, $d \log \pi_\theta \propto (1 + \gamma_\theta - \rho_\theta) \frac{d \log s_\theta}{d\theta}$, and s_θ is increasing in θ.

s_θ, and firm types θ :

$$\frac{d\log\epsilon_\theta}{d\theta} = \frac{\mu_\theta - \epsilon_\theta}{\epsilon_\theta}\frac{d\log s_\theta}{d\theta}. \tag{40}$$

Given estimates of markups across the sales distribution, this differential equation informs ϵ_θ up to a boundary condition ϵ_{θ^*}, which is chosen to match a given value of average consumer rents $\mathbb{E}_s\left[\epsilon_\theta\right] = \bar{\epsilon}$.

Similarly, worker rents δ_θ vary across firms of different types as follows

$$\frac{d\log\delta_\theta}{d\theta} = \mu_\theta(\frac{1}{\delta_\theta} - \frac{1}{\mathcal{M}_\theta})\frac{d\log s_\theta}{d\theta}. \tag{41}$$

Given markdowns and sales estimates, worker rents δ_θ cup to a boundary condition δ_{θ^*}. Conditioning on θ^* and normalizing $L = M = 1$, the free entry and selection conditions inform the labor cost of entry and overhead production, $w_o f_o$ and $w_e f_e$,

$$f_e w_e + (1 - G(\theta^*))w_o f_o = \mathbb{E}_s\left[1 - \frac{\mathcal{M}_\theta}{\mu_\theta}\right],$$

$$(1 - G(\theta^*))w_o f_o = s_{\theta^*}(1 - \frac{\mathcal{M}_{\theta^*}}{\mu_{\theta^*}}).$$

To inform $\{\mathcal{M}_i, \delta_i\}_{i\in\{o,e\}}$, I utilize the following first-order relationships,

$$\log\delta_i/\delta_\theta = \mu_\theta(\frac{1}{\delta_\theta} - \frac{1}{\mathcal{M}_\theta})\log\frac{w_i f_i}{\omega_\theta^{wn}}. \tag{42}$$

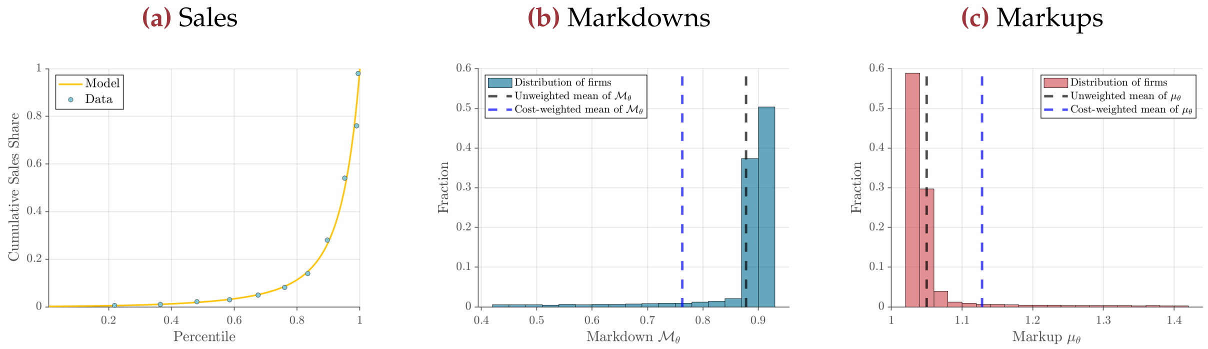
Given a boundary condition for δ_{θ^*}, (40) pins down $\{\delta_\theta\}_{\theta\in\Theta}$. Choosing firm types θ with comparable wage bills, (42) informs workers' rents in entry and overhead producing jobs, δ_e and δ_o. Given worker rents in the entry and overhead sectors, I approximate $\mathcal{M}_o$ and $\mathcal{M}_e$ by $\log\frac{\delta_o}{\delta_e} = \frac{1}{\mathcal{M}_e}\frac{\mathcal{M}_e-\delta_e}{\delta_e}\log\frac{w_o f_o}{w_e f_e}$ and $\log\frac{\delta_e}{\delta_0} = \frac{1}{\mathcal{M}_o}\frac{\mathcal{M}_o-\delta_o}{\delta_o}\log\frac{w_o f_o}{w_e f_e}$.

Finally, I calibrate θ^* to match standard estimates of firm exit hazard rates.

6.2 Calibration Results

The calibration approach requires information on markdowns, markups, and sales shares of final good firms $\{\mathcal{M}_\theta, \mu_\theta.s_\theta\}$. A number of papers proivde joint estimates of markups and markdowns across establishments, e.g., in the U.S. (**?**) and Germany (Dolfen, 2020), utilizing the the production function approach outlined in Loecker & Warzynski (2012). Both find that markups and markdowns positively covary with firm employment, and sales. For the baseline, I construct markups and markdowns across the sales distribution consistent with these state-of-the-art estimates. Appendix C provides further details on the implementation.

Figure 2 Distribution of Sales, Markdowns and Markups


(a) Sales **(b)** Markdowns **(c)** Markups

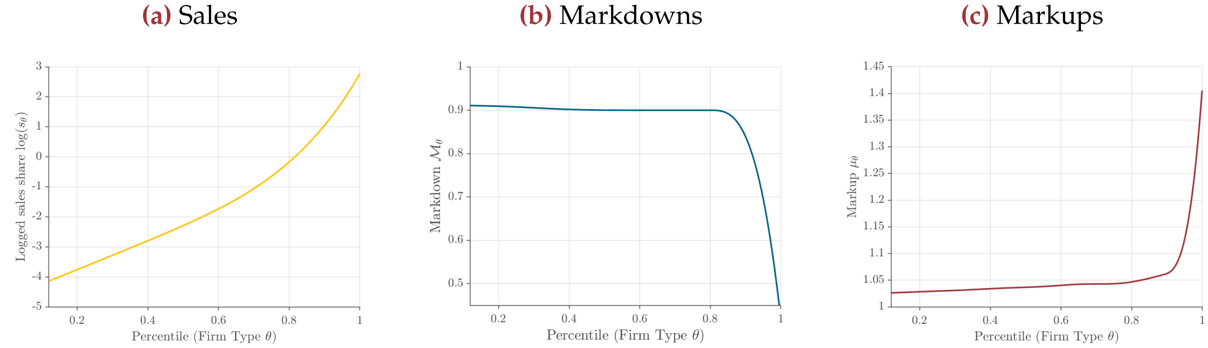
Notes: This figure plots the calibrated distribution of sales in panel (a), of markdowns in panel (b), and of markups in panel (c).

For the baseline calibration, I choose boundary conditions that imply efficient selection. Following Lemma 3, I impose $\epsilon_{\theta^*} = \bar{\epsilon}$ and $\frac{\mathcal{M}_{\theta^*}}{\mu_{\theta^*}}\delta_{\theta^*} + (1 - \frac{\mathcal{M}_{\theta^*}}{\mu_{\theta^*}})\delta_o = \bar{\delta}$. Further, I choose θ^* to match a firm exit rate of 10%.

Figure 2 displays the calibrated distributions of sales, markdowns, and markups. Both the distribution of markdowns and markups display significant amounts of dispersion. The unweighted mean of markdowns equals 12 percent, the cost-weighted average markdown equals 23 percent. Similarly, the unweighted mean of markups equals 4%, while the cost-weighted mean of markups equals 15 percent.

Figure 3 plots productivity $\log \tilde{A}_\theta = \log A_\theta A_\theta^n A_\theta^c$, markdowns, and markups by firm type θ. Markdowns and markups vary relatively little within the 80th percentile of the productivity distribution, but increase steeply thereafter.

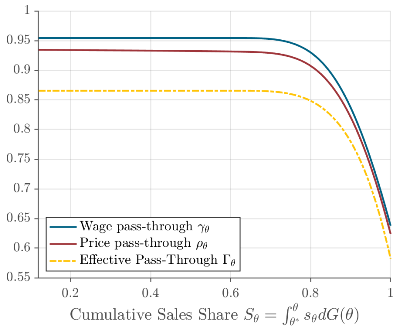
Figure 3 Sales, Markdowns, and Markups by Firm Type


(a) Sales **(b)** Markdowns **(c)** Markups

Notes: This figure plots sales, markdowns, and markups as a function of firm type in the calibrated market allocation.

Figure 4 displays how average firms' price and cost pass-throughs vary across the sales distribution. Consistent with facts documented in the literature, pass-throughs are nearly complete within the first four quintiles of the sales distributuon, but fall sharply in the top quintile. [30]The figure also shows that the structural pass-through of "productivity shocks" into prices implied by the model is substantially lower, capturing the double-marginalization of prices.

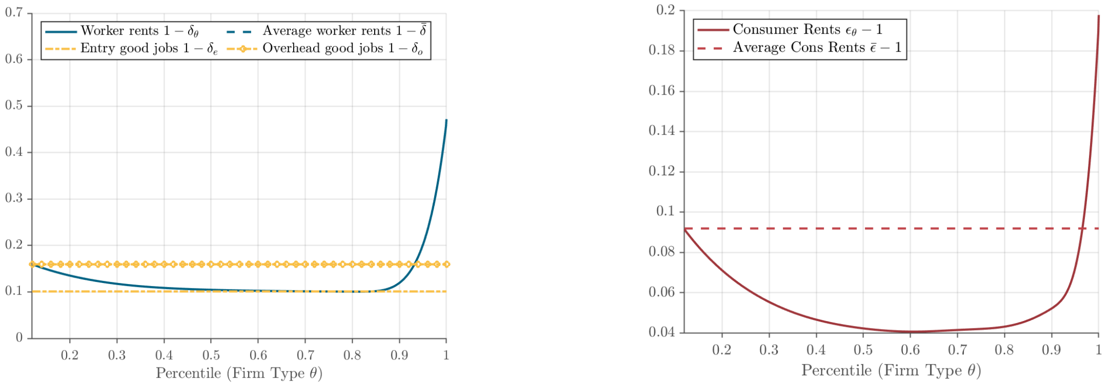
Figure 4 Calibrated Pass-Through by Firm type



Notes: This figure plots the pass-through of cost shocks into prices in red, the pass-through of labor productivity shocks into wages in blue, and the effective price pass-through, $\frac{\partial \log p_\theta}{\partial \log \tilde{A}_\theta}$,as a function of firm type θ.

Finally, figure 5 shows that rents in labor and product markets are U-shaped in effective productivity. Firms that earn the highest rents in labor and product markets also provide employees and costumers with high worker and consumer rents. The total surplus earned by households, $\mathbb{E}_s [\epsilon_\theta] / \mathbb{E}_{wn} [\delta_\omega] - 1$, equals 29 percent, while the harmonic sales-weighted average of firm rents equals 37 percent. Following Lemma 2, entry is excessive in the initial allocation.

Figure 5 Worker and Consumer Rents by Firm Type


(a) Worker Rents

(b) Consumer Rents

Notes: This figure plots the rents earned by households in in labor and product markets by producer type.

[30] The calibrated price pass-throughs are consistent with the patterns documented by Amiti *et al.* (2019): While firms with low sales have near complete pass-through, pass-throughs decline sharply in the upper quintile of the firm sales distribution.

6.3 Quantitative Welfare Analysis

6.3.1 Total Welfare Loss from Imperfect Competition

Using the calibrated firm-level elasticities, I use Proposition 2, Proposition 3, and Theorem 2 to calculate the societal costs of distortions in consumption equivalent terms. Table 1

Table 1 The Costs of Imperfect Competition

C.E. Welfare Loss			14.5%	
Allocative Losses			10.5%	
Deadweight Loss			3.25%	
Effect	Direct Efficiency		Factor Supply	
Contribution	37.8%		62.2%	
Effect	Indirect Allocative		Indirect Allocative	DWL
	Firm size	Entry		
Contribution	33.3%	6.2%	36.5%	24.0%

Notes: This table displays the consumption equivalent welfare loss implied by Proposition 2, Theorem 2, and Proposition 3. The Frisch elasticity of factor supply is set to $\varphi = 0.5$.

The consumption equivalent welfare loss equals 13 percent. Around 75 percent of the costs arise from allocative inefficiencies, implying that micro-level heterogeneity in markdowns and markups account for the bulk of the total welfare loss. Holding fixed the supply of factors, factor misallocations account for 40 percent of the total welfare loss. This shows that indirect efficiency losses compound the costs of factor misallocations. A key implication is that entry is an effective tool to remove distortions from imperfect competition in this economy.

6.3.2 The Role of Factor and Product Markets as Catalysts of Indirect Efficiency Losses

As discussed previously, factor supply distortions may amplify or alleviate allocative efficiency losses. Following Theorem 2, indirect efficiency losses can be decomposed into costs catalyzed by factor and product markets.The following table breaks down the contribution of each of these components to the total productivity loss.

Table 2 Indirect Allocative Efficiency Losses

Contribution to Total	Product Market Competition	Factor Market Competition	Interaction
Total	61.6%	-4.8%	42.9%
$\zeta^{\cong}_{\Sigma}, \zeta_{\boldsymbol{\Sigma}} + \zeta_{\mathcal{E}}$	44%		49%
$\zeta_{\boldsymbol{\Gamma}}$	15%		-17%

Notes: This table decomposes the allocative efficiency losses implied by Proposition 3 into the effects of labor and product markets. The elasticity of factor supply is set to $\varphi = 0.5$.

Interactions between markups and markdowns catalyze over a third of the indirect efficiency losses from market power. In both product and labor markets, adjustments in entry constitute the dominant source of efficiency losses.[31] The decomposition reveals that competition in factor markets implies greater overall efficiency gains through this margin. Importantly, entry adjustments in factor markets trigger an additional effect that operates through quasi-rents.

In contrast, factor market competition is "anti-competitive," implying that adjustments in markups and markdowns in response to a change in the mass of entrants lower welfare. Hence, this dampens indirect efficiency losses. In contrast, an increase in the mass of entrants has strong pro-competitive effects in product markets. This provides an example of how an increase in competition can trigger reallocations that are beneficial in one, but cause efficiency losses in another market, underlining the importance of separately accounting for factor and product markets.

6.4 Summary

The quantitative welfare results suggest that micro-level heterogeneity in markdowns and markups accounts for the bulk of the societal costs of market power. Compared to an economy with comparable firm-level and aggregate distortions but perfectly competitive labor markets, monopsony not only increases overall welfare losses, but also raises the relative importance of allocative inefficiencies. Insufficient scale, in turn, accounts for a large share of the overall allocative inefficiencies. These findings have important implications for the design of policy aimed at addressing inefficiencies from market power, which I address in the next section.

[31] This is relates closely to recent work by Baqaee *et al.* (2022). Calling this the Darwinian effect, they show that it drives most of the allocative efficiency gains from an increase in market size in an economy with heterogeneous markups.

7 Extensions

This section discusses extensions of the baseline framework along multiple dimensions. Details are relegated to Appendix D. First, I briefly how to integrate local labor markets into the model. Second, I extend the model to account for heterogeneity in worker types, e.g., skill or occupations. Third, I consider an alternative labor supply systems that also features variable labor supply elasticities.

HSA and VES type labor supply Appendix D.2 formulates alternative versions of the model where households are endowed with HSA (Matsuyama & Ushchev (2017)) and VES-type preferences over jobs and varieties. Theorem 1 and Proposition 1 continue to apply in this setting. However, VES-type preferences, due to their non-homotheticity, effects channel additional welfare effects that are unrelated to technical and allocative efficiency.

Heterogeneous worker groups Appendix D.3 extends the model to allow for heterogeneous worker groups, for example heterogeneity in skill or occupations, and derives necessary and sufficient conditions, analogous to Theorem 1, for the resulting market equilibrium to induce efficient allocations, conditional on factor supply.

8 Conclusion

A growing body of empirical evidence highlights the prevalence of imperfect competition in both labor and product markets, raising concerns about its implications for welfare and inequality. This paper proposes a new framework suited to analyze misallocations caused by imperfect competition in labor markets in the presence of firm heterogeneity, imperfect competition in product markets, and fixed costs.

Throughout, this paper emphasizes the idea that the joint distribution of labor and product market rents is key not only to assess welfare losses from imperfect competition quantitatively, but also qualitatively. My theoretical results provide sharp characterizations of the inefficiencies posed by imperfect competition in both labor and product markets. I show that allocative inefficiencies materialize not only through dispersion in rents that are observable, but also through potentially foregone allocative gains from scale. I show theoretically and quantitatively that this distinction is key for how welfare-enhancing policies can be designed in light of the substantial heterogeneity in wage markdowns and price markups documented by previous work. Further, my results show that the interaction of labor and product market power gives rise to quantitatively important margins of allocative inefficiency absent in prior theoretical and empirical work.

The contributions of this paper provide exciting avenues for future research. For example, I abstract from the local nature of labor markets or the distributional consequences of imperfect competition in labor markets. Due to its tractability, the model provides a natural starting point to address, e.g., the effects of local competition policies, minimum wages, or progressive income taxation in an economy with imperfect competition in both input and output markets.

References

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

Arkolakis, Costas, Arnaud Costinot, Dave Donaldson and Andrés Rodrìguez-Clare, 'The Elusive Pro-Competitive Effects of Trade.' *Review of Economic Studies*, **86** (1), pp. 46–80, *2019*.

_ , _ and Andres Rodriguez-Clare, 'New Trade Models, Same Old Gains?.' *American Economic Review*, **102** (1), pp. 94–130, *2012*.

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

Baqaee, David, Emmanuel Fahri and Kunal Sangani, 'The Darwinian Returns to Scale.' working paper, *2022*.

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

Behrens, Kristian, Giordano Mion, Yasusada Murata and Jens Suedekum, 'Quantifying the Gap Between Equilibrium and Optimum under Monopolistic Competition*.' *The Quarterly Journal of Economics*, **135** (4), pp. 2299–2360, *2020*, DOI: `http://dx.doi.org/10.1093/qje/qjaa017`.

Berger, David W., Kyle F. Herkenhoff and Simon Mongey, 'Labor Market Power.' *American Economic Review*, 2022.

Bilbiie, Florin O., Fabio Ghironi and Marc J. Melitz, 'Monopoly Power and Endogenous Product Variety: Distortions and Remedies.' *American Economic Journal: Macroeconomics*, **11** (4), pp. 140–74, *2019*, DOI: `http://dx.doi.org/10.1257/mac.20170303`.

Brooks, Wyatt J., Joseph P. Kaboski, Yao Amber Li and Wei Qian, 'Exploitation of labor? Classical monopsony power and labor's share.' *Journal of Development Economics*, **150** (C), *2021*, DOI: `http://dx.doi.org/10.1016/j.jdeveco.2021.10`.

Card, David, Ana Rute Cardoso, Joerg Heining and Patrick Kline, 'Firms and Labor Market Inequality: Evidence and Some Theory.' *Journal of Labor Economics*, **36** (S1), pp. 13–70, *2018*.

Chan, Mons, Sergio Salgado and Ming Xu, 'Heterogeneous Passthrough from TFP to Wages.' working paper, *2021*.

— **, Ming Xu and Sergio Salgado**, 'Heterogeneous Passthrough from TFP to Wages.' 2019 Meeting Papers 1447, Society for Economic Dynamics, *2019*.

Dhingra, Swati and John Morrow, 'Monopolistic Competition and Optimum Product Diversity under Firm Heterogeneity.' *Journal of Political Economy*, **127** (1), pp. 196 – 232, *2019*.

Dixit, Avinash K and Joseph E Stiglitz, 'Monopolistic Competition and Optimum Product Diversity.' *American Economic Review*, **67** (3), pp. 297–308, *1977*.

Dolfen, Paul, 'The Decline of the Labor Share: Markups, Markdowns, or Technology?' working paper, *2020*.

Dube, Arindrajit, Jeff Jacobs, Suresh Naidu and Siddharth Suri, 'Monopsony in Online Labor Markets.' *American Economic Review: Insights*, **2** (1), pp. 33–46, *2020*, DOI: `http://dx.doi.org/10.1257/aeri.20180150`.

Edmond, Chris, Virgiliu Midrigan and Daniel Yi Xu, 'How Costly Are Markups?.' NBER Working Papers 24800, National Bureau of Economic Research, Inc, *2021*.

Epifani, Paolo and Gino Gancia, 'Trade, markup heterogeneity and misallocations.' *Journal of International Economics*, **83** (1), pp. 1–13, *2011*.

Garin, Andrew and Filipe Silvero, 'How Responsive are Wages to Demand within the Firm? Evidence from Idiosyncratic Export Demand Shocks.' working paper, *2018*.

Haanwinckel, Daniel, 'Supply, Demand, Institutions and FIrms: A Theory of Labor Market Sorting and the Wage Distribution.' working paper, *2021*.

Hall, Robert, 'The Relation between Price and Marginal Cost in U.S. Industry.' *Journal of Political Economy*, **96** (5), pp. 921–47, *1988*.

Jha, Priyaranjan and Antonio Rodriguez-Lopez, 'Monopsonistic labor markets and international trade.' *European Economic Review*, **140** (C), *2021*, DOI: `http://dx.doi.org/10.1016/j.euroecorev.2021`.

Keane, Michael and Richard Rogerson, 'Micro and Macro Labor Supply Elasticities: A Reassessment of Conventional Wisdom.' *Journal of Economic Literature*, **50** (2), pp. 464–76, *2012*, DOI: `http://dx.doi.org/10.1257/jel.50.2.464`.

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

Kroft, Kory, Yao Luo, Magne Mogstad and Bradley Setzler, 'Imperfect Competition and Rents in Labor and Product Markets: The Case of the Construction Industry.' Working Papers tecipa-666, University of Toronto, Department of Economics, *2020*.

Krugman, Paul R., 'Increasing returns, monopolistic competition, and international trade.' *Journal of International Economics*, **9** (4), pp. 469–479, *1979*.

Lamadon, Thibaut, Magne Mogstad and Bradley Setzler, 'Imperfect Competition, Compensating Differentials, and Rent Sharing in the US Labor Market.' *American Economic Review*, **112** (1), pp. 169–212, *2022*, DOI: `http://dx.doi.org/10.1257/aer.20190790`.

Lerner, A. P., 'The Concept of Monopoly and the Measurement of Monopoly Power.' *Review of Economic Studies*, **1** (3), pp. 157–175, *1934*.

Loecker, Jan De and Frederic Warzynski, 'Markups and Firm-Level Export Status.' *American Economic Review*, **102** (6), pp. 2437–2471, *2012*.

Mankiw, N. Gregory and Michael D. Whinston, 'Free Entry and Social Inefficiency.' *RAND Journal of Economics*, **17** (1), pp. 48–58, *1986*.

Manning, Alan, 'The real thin theory: monopsony in modern labour markets.' *Labour Economics*, **10** (2), pp. 105–131, *2003*.

— , 'Monopsony in Labor Markets: A Review.' *ILR Review*, **74** (1), pp. 3–26, *2021*, DOI: `http://dx.doi.org/10.1177/0019793920922499`.

Matsuyama, Kiminori and Philip Ushchev, 'Beyond CES: Three Alternative Classes of Flexible Homothetic Demand Systems.' CEPR Discussion Papers 12210, C.E.P.R. Discussion Papers, *2017*.

— **and** — , 'When Does Procompetitive Entry Imply Excessive Entry?.' CEPR Discussion Papers 14991, C.E.P.R. Discussion Papers, *2020*.

— **and** — , 'Selection and Sorting of Heterogeneous Firms through Competitive Pressures.' working paper, *2022*.

Melitz, Marc J., 'The impact of trade on intra-industry reallocations and aggregate industry productivity.' *Econometrica*, **71** (6), pp. 1695–1725, *2003*.

— **and Gianmarco I. P. Ottaviano**, 'Market Size, Trade, and Productivity.' *Review of Economic Studies*, **75** (1), pp. 295–316, *2008*.

— **and Stephen J. Redding**, 'New Trade Models, New Welfare Implications.' *American Economic Review*, **105** (3), pp. 1105–1146, *2015*.

Mrázová, Monika and J. Peter Neary, 'Not So Demanding: Demand Structure and Firm Behavior.' *American Economic Review*, **107** (12), pp. 3835–3874, *2017*.

— **and** — , 'Selection Effects with Heterogeneous Firms.' *Journal of the European Economic Association*, **17** (4), pp. 1294–1334, *2019*.

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

Rosen, Sherwin, 'The theory of equalizing differences.' In O. Ashenfelter and R. Layard (eds.) *Handbook of Labor Economics*, **1** of Handbook of Labor Economics: Elsevier, Chapter 12, pp. 641–692, *1987*.

Samuelson, Paul A, *Foundations of Economic Analysis*: Harvard University Press, *1947*.

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

Spence, A., 'Product Selection, Fixed Costs, and Monopolistic Competition.' *Review of Economic Studies*, **43** (2), pp. 217–235, *1976*.

Staiger, Doug, Joanne Spetz and Ciaran S. Phibbs, 'Is There Monopsony in the Labor Market? Evidence from a Natural Experiment.' *Journal of Labor Economics*, **28** (2), pp. 211–236, *2010*.

Thisse, Jaques-Francois and Philip Ushchev, 'When can a demand system, be described by a multinomial logit with income effect?' working paper, *2016*.

Trottner, Fabian, 'Who gains from scale? Trade and Wage Inequality between firms.' working paper, *2020*.

Venables, Anthony, 'Trade and trade policy with imperfect competition: The case of identical products and free entry.' *Journal of International Economics*, **19** (1-2), pp. 1–19, *1985*.

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

Yeh, Chen, Claudia Macaluso and Brad Hershbein, 'Monopsony in the US Labor Market.' *American Economic Review*, **112** (7), pp. 2099–2138, *2022*.

Zhelobodko, Evgeny, Sergey Kokovin, Mathieu Parenti and Jacques-François Thisse, 'Monopolistic Competition: Beyond the Constant Elasticity of Substitution.' *Econometrica*, **80** (6), pp. 2765–2784, *2012*, DOI: `http://dx.doi.org/ECTA9986`.

Appendix

A Derivations

A.1 Household Problem

Households maximzie utility choosing how many hours to supply to firms and how much to consume:

$$\max_{C,N,c_\theta,n_\omega} U(C,N)$$

subject to the the following constraints:

$$1 = \int \Upsilon_\theta(\frac{c_\theta}{C})dM^C(\theta)$$

$$1 = \int \Psi_\omega\left(\frac{n_\omega}{N}\right)dM^E(\omega)$$

$$\int p_\theta c_\theta dM^C(\theta) = \int n_\omega w_\omega dM^E(\omega)$$

Denoting the multipliers of the dual problem for the associated constraints by λ_C, λ_N, and γ, the first order conditions with respect to C, N, c_ω and $n_{\omega'}$ are given by:

$$U_c C = -\lambda_C \int \Upsilon'_\theta(\frac{c_\theta}{C})\frac{c_\theta}{C}dM^C(\theta) \tag{A.1}$$

$$U_N N = \lambda_N \int \Psi'_\omega(\frac{n_\omega}{N})\frac{n_\omega}{N}dM^E(\omega) \tag{A.2}$$

$$-\lambda_C \Upsilon'_\theta(\frac{c_\theta}{C})\frac{1}{C} = \gamma p_\theta \tag{A.3}$$

$$\lambda_N \Psi'_\omega(\frac{n_\omega}{N})\frac{1}{N} = \gamma w_{\omega'} \tag{A.4}$$

Using (A.1) to substitute λ_C in (A.3) yields:

$$\frac{U_c C}{\int \Upsilon'_\theta(\frac{c_\theta}{C})\frac{c_\theta}{C}dM^C(\theta)}\Upsilon'_\theta(\frac{c_\theta}{C})\frac{1}{C} = \gamma p_\theta$$

Multiplying both sides by c_θ, integrating over all consumption varieties and plugging into the budget constraint, we obtain:

$$\frac{U_C C}{Y} = \gamma$$

Defining $\mathcal{P} = \frac{1}{C \int \Upsilon'_\theta(\frac{c_\theta}{C})\frac{c_\theta}{C} dM^C(\theta)}$, the demand for variety ω can be written:

$$\frac{p_\theta}{\mathcal{P}} = \Upsilon'_\theta(\frac{c_\theta}{C}). \tag{A.5}$$

Analogous derivations imply that $-\frac{U_N N}{Y} = \gamma$, and for $\mathcal{W} = \frac{1}{N \int \Psi'(\frac{n_\omega}{N})\frac{n_\omega}{N} dM^E(\omega)}$, labor supply to employer ω is given by:

$$\frac{w_\omega}{\mathcal{W}} = \Psi'_\omega(\frac{n_\omega}{N}). \tag{A.6}$$

A.2 Microfoundation of the Labor Supply System

I develop a model of random dicrete choice that microfounds the factor supply system in the main text. I describe the problem of workers that choose how one of many jobs $\omega \in \Omega$ to meet an income target $y_{i.}$

There is a continuum of workers i of mass L. Workers optimally pick an firm ω. Preferences for leisure and consumption are separable, so the problem of choosing an employer can be analyzed supposing that a worker i has to earn some level of income $y_i \sim F(y)$. Workers provide $n_{i,\omega} = y_i/w_\omega$ hours of work to a firm ω offering a wage w_ω.

The indirect disutility for a worker that has to earn income y_i and chooses to work for firm ω when faced with a schedule of wage offers $\{w_{\omega'}\}_{\omega' \in \Omega'}$, is assumed to take the following form:

$$V_{\omega i} = \mu(\ln \left[\frac{w_\omega}{\mathcal{W}} \left(\Psi'_\omega\right)^{-1} \left(\frac{w_\omega}{\mathcal{W}}\right)\right] - \ln y_i) - \epsilon_{\omega i},$$

where $\mathcal{W}$ is a wage index solving $\int \Psi_\omega \left(\left(\Psi'_\omega\right)^{-1} \left(\frac{w_{\omega'}}{\mathcal{W}}\right)\right) d\omega = 1$, and $\epsilon_{\omega i}$ is an idiosyncratic preference shock that is i.i.d. Gumbel distributed with standard deviation $\mu\pi/\sqrt{6}$. $\Psi(.)$ is a strictly increasing, convex function. It is worth noting that the indirect utility corrseponds to the canonical model of multinomial discrete choice that microfounds CES-type labor supply systems in the special case where $\Psi_\omega(x) = a_\omega x^{\frac{\beta+1}{\beta}}$.

The key departure from the standard multinomial choice framework is that the relative disutility received from working for two different employers depends on the whole set of possible alternatives rather than only the wage and non-wage amenities offered by the two firms. The preferences above thus imply a departure from the Independence of Irrelevant Alternatives property that is inherent to CES utility. A natural interpretation is that the perceived utility from different options is influenced by the menu from which this choice is made (Sen, 1997). Departing from the IIA property provides the multinomial discrete choice model with sufficient flexibility to microfound the aggregate labor supply system from the main text.

The probability that an individual optimally chooses to work for employer ω is independent of income y:

$$\pi_\omega = \frac{\frac{w_\omega}{\mathcal{W}}(\Psi'_\omega)^{-1}(\frac{w_\omega}{\mathcal{W}})}{\int \frac{w_{\omega'}}{\mathcal{W}}\left(\Psi'_{\omega'}\right)^{-1}\left(\frac{w_{\omega'}}{\mathcal{W}}\right)d\omega'},$$

By the LLN, this probability will also equal the share of workers that choose to work for employer ω. Total expected hours supplied by worker i to firm ω equal:

$$n_{\omega,i} = \frac{y_i}{w_i}\pi_\omega = y_i \frac{(\Psi'_\omega)^{-1}(\frac{\omega}{\mathcal{W}})}{\mathcal{W}\int \frac{w_{\omega'}}{\mathcal{W}}\left(\Psi'_{\omega'}\right)^{-1}\left(\frac{w_{\omega'}}{\mathcal{W}}\right)d\omega'}$$

The average per-capita labor supply of hours to firm ω is given by:

$$n_\omega \equiv \int_i n_{\omega,i}di = \frac{(\Psi'_\omega)^{-1}(\frac{w_\omega}{\mathcal{W}})}{\mathcal{W}\int \frac{w_{\omega'}}{\mathcal{W}}\left(\Psi'_{\omega'}\right)^{-1}\left(\frac{w_{\omega'}}{\mathcal{W}}\right)}\int_i y_i dF(y_i).$$

Noting that the integral equals per-capita nominal GDP, since $\int_i y_i dF(y_i) \equiv Y$. Noting that $Y\mathcal{W}\int \frac{w_{\omega'}}{\mathcal{W}}\left(\Psi'_{\omega'}\right)^{-1}\left(\frac{w_{\omega'}}{\mathcal{W}}\right) = N$, we recover the demand system used in the main text.

A.3 Alternative Model formulation for Overhead Inputs

Here I show how to relax the assumptions that overhead inputs are produced outside and homogeneous across firms. To do so, we can express the labor supply index N as:

$$1 = M\Psi_e(\frac{n}{N}) + M\int_{\theta^*}^{\infty}(\Psi_\theta(\frac{n_\theta}{N}) + \Psi_{\theta,o}(\frac{n_{\theta,o}}{N}))dG(\theta), \tag{A.7}$$

where n_θ denotes the jobs in variable production at firm θ, while $n_{\theta,o}$ denotes labor allocated to the production of overhead inputs at θ. Importantly, from the perspective of workers, these jobs are differentiated. This formulations has the advantage that it tractably allows for endogenous overhead costs that also vary across firms. I assume that employees perform jobs that they were hired for. That is, a worker hired to produce overhead inputs cannot work in variable good production, and vice versa.

The output prices and wages offered to workers in variable production are determined by the same equations as in the main text. Firms offer wage to employees in overhead jobs that satisfy: $w_{\theta,o} = \mathcal{W}\Psi'_{\theta,o}(\frac{f_{\theta,o}}{LN})$, where $f_{\theta,o}$ is the quantity of overhead inputs required for production for a firm of type θ. If $\Psi_{\theta,o}(x) = \Psi_o(x)$ and $f_{\theta,o} = f_o$, one recovers the model in the main text.

A firm decides to produce if, and only, if:

$$\frac{(1-\frac{\mathcal{M}_\theta}{\mu_\theta})p_\theta c_\theta}{w_{\theta,o}f_{\theta,o}} \geq 1/L. \tag{A.8}$$

The existence of a unique selection cutoff θ^* requires that final good firms can be ordered so that the left-hand-side is strictly increasing and monotonous in θ. The free entry condition can then be expressed as,

$$\int_{\theta^*}^{\infty}((1-\frac{\mathcal{M}_\theta}{\mu_\theta})p_\theta c_\theta - w_{\theta,o}f_{\theta,o})dG(\theta) = p_e f_e,$$

where p_e is determined by the same equation as in the main text.

The results regarding the efficiency of the market allocation in Section 3 remain unchanged under this alternative model formulation. All welfare results continue to hold, but $\overline{\mathcal{M}}_f = \mathbb{E}_{wn}\left[\frac{1}{\mathcal{M}_x}|x \in \{(\theta,o)_\Theta, e\right]$.

A.4 Example of an Efficiency-Inducing Policy

The following provides an of a set of policies τ^* that implements the efficient allocation through taxes and subsidies. $\tau^* = \{\tau_\theta^{pc}, \tau_\theta^{wn}, \tau_e, \tau_o\}$ is given by,

$$\begin{array}{ll} \text{Sales subsidy} & \tau_\theta^{pc}(\frac{c}{C}) = [\Upsilon_\theta(\frac{c}{C}) - \Upsilon'_\theta(\frac{c}{C})]\bar{P}C, \\ \text{Wage-bill subsidy} & \tau_\theta^{wn}(\frac{n}{N}) = [\Psi_\theta\left(\frac{n}{N}\right) - \Psi'_\theta(\frac{n}{N})]\bar{W}N, \\ \text{Entry tax} & \tau_e(p_e f_e) = \left(\delta_e/\bar{W} - 1\right)p_e f_e \bar{W}N, \\ \text{Overhead tax} & (p_o f_o) = (\delta_o/\bar{W} - 1)p_e f_e \bar{W}N, \end{array} \tag{A.9}$$

Policy τ^* ensures that equilibrium prices and wages are set competitively. To preserve entry incentives, subsidies ensure that private profits of firms equal the social profit they generate, $\pi_\theta^{\text{opt}}/w_\theta^{\text{opt}}n_\theta^{\text{opt}} = \frac{CS_\theta^{\text{opt}}+WS_\theta^{\text{opt}}}{\delta_\theta^{\text{opt}}}$. To achieve this, marginal sales and wage bill subsidies equal firms' desired markups and markdowns respectively, $t'_\theta(pc) = \mu_\theta(pc)$ and $\tau'_\theta(wn) = \mathcal{M}_\theta(wn)$. Marginal subsidies are increasing in the market power a firm excerts in product and labor markets, while the total amount of subsidies received relative to a firm's sales depends on the worker and consumer rents it generates.

The production of entry and overhead goods may be taxed or subsidized. Since $\bar{W} = \mathbb{E}_{wn}[\delta]$, entry and overhead goods ought to be taxed if they provide lower worker surplus than the average job in the economy, and vice versa. Intuitively, only by ensuring that worker surpluses generated through expansionary activities are socially optimal, the policy can induce efficient entry and selection.

B Proofs

Proof of Theorem 1

Proof. The proof assumes that an equilibrium exists and is unique. See Appendix B for a discussion of the equilibrium properties.

To prove the "if" part, I show that the conditions pinning down the planner's allocation coinicde with those that determine the market allocation under the conditions stated in the theorem, which are equivalent to requiring that $\forall \theta \in$ support$\{G(\theta)\}$, $\epsilon_\theta \equiv \mu = \frac{\sigma}{\sigma-1}$ and $\forall \omega' \in \{o, e, \{\theta\}_{\theta \in \text{support}\{G(\theta)\}}\}$, $\delta_{\omega'} \equiv \mathcal{M} = \frac{\beta}{\beta+1}$.

First, I rewrite the planner's problem. Fixed costs imply that the planner chooses a cutoff θ^* such that variable production equals zero for varieties with draws $\theta < \theta^*$. Second, convexity of $\Psi(.)$ implies that the planner optimally allocates $n_o = \frac{f_o}{L}$ and $n_e = \frac{f_e}{L}$ workers to the production of entry and overhead goods. The problem of the planner can, this, be written:

$$\begin{aligned} \mathcal{L} = \max_{C,,c_\theta,,M_e,\theta^*,\lambda_C,\lambda_N} \quad & U(C, \bar{N}) + \lambda_C \left[1 - M_e \textstyle\int_{\theta^*}^{\infty} \Upsilon(\frac{c_\theta}{C}) dG(\theta)\right] \\ & + \lambda_N \left[M_e(\Psi_e(\tfrac{f_e}{\bar{N}L}) + \textstyle\int_{\theta^*}^{\infty}(\Psi(\frac{c_\theta/A_\theta}{\bar{N}}) + \Psi_o(\frac{f_o}{\bar{N}L}))dG(\theta)) - 1\right] \end{aligned}$$

Following the main text, denote $\epsilon_\theta \equiv \frac{\Upsilon_\theta(\frac{c_\theta}{C})}{\Upsilon'_\theta(\frac{c_\theta}{C})\frac{c_\theta}{C}}$ and $\delta_{\omega'} = \frac{\Psi_{\omega'}(\frac{n_{\omega'}}{N})}{\Psi'_{\omega'}(\frac{n_{\omega'}}{N})\frac{n_{\omega'}}{N}}$. The planner's first order condition with respect to c_θ can be written:

$$\Upsilon(\frac{c_\theta}{C}) = \frac{\epsilon_\theta}{\delta_\theta}(\frac{\lambda_N}{\lambda_C})\Psi(\frac{c_\theta}{\bar{N}A_\theta}). \tag{B.1}$$

The first order condition with respect to M_e implies:

$$\frac{\lambda_N}{\lambda_C} = \frac{\int_{\theta^*}^{\infty} \Upsilon(\frac{c_\theta}{C}) dG(\theta)}{\Psi(f_e/\left(L\bar{N}\right)) + \int_{\theta^*}^{\infty} \left\{\Psi(f_o/(L\bar{N})) + \Psi(\frac{n_\theta}{\bar{N}})\right\} dG(\theta)} = 1, \tag{B.2}$$

where the last equality follows imposing that all constraints bind.

The planner's first order conditions with respect to C reads

$$U_C C = -\lambda_c M \int_{\theta^*}^{\infty} \Upsilon'(\frac{c_\theta}{C})\frac{c_\theta}{C} dG(\theta) \tag{B.3}$$

Finally, the planners FOC pinning down the selection cutoff is given by:

$$\lambda_C \Upsilon(\frac{c_{\theta^*}}{C}) - \lambda_N \Psi(\frac{c_{\theta^*}}{\bar{N}A_{\theta^*}}) = \lambda_N \Psi\left(\frac{f_o}{NL}\right) \tag{B.4}$$

I now show that if $\forall \theta\ \epsilon_\theta \equiv \mu = \frac{\sigma}{\sigma-1}$, and $\forall \omega'$, $\delta_{\omega'} \equiv \mathcal{M} = \frac{\beta}{\beta+1}$, then the planner chooses the same labor allocations across firms, selection cutoff, and aggregate consumption index C as the market. First, note that profit-maximization of firms implies that wages and prices are related through:

$$p_\theta = \frac{\mu_\theta}{\mathcal{M}_\theta} \frac{w_\theta}{A_\theta}. \tag{B.5}$$

When $\epsilon_\theta \equiv \mu = \frac{\sigma}{\sigma-1}$ and $\delta_{\omega'} \equiv \mathcal{M} = \frac{\beta}{\beta+1}$, then $\mathcal{P}$ in (6) and $\mathcal{W}$ in (7) can be expressed as $\mathcal{P} = \frac{1}{C}\mu$ and $\mathcal{W} = \frac{1}{\bar{N}}\mathcal{M}$. In this case, we can rewrite per-capita labor supply in (5) and product demand in (4) as $\mu \Upsilon'(\frac{c_\theta}{C})\frac{1}{C}Y = p_\theta$ and $\mathcal{M}\Psi'(\frac{n_\theta}{\bar{N}})\frac{1}{\bar{N}}Y = w_\theta$. Substituting those expressions into B.5 and imposing $\mu_\theta = \mu$ and $\mathcal{M}_\theta = \mathcal{M}$, we obtain $\Upsilon'(\frac{c_\theta}{C})\frac{1}{C} = \Psi'(\frac{c_\theta}{\bar{N}A_\theta})\frac{1}{A_\theta \bar{N}}$. Multiplying both sides by c_θ, when $\epsilon_\theta \equiv \mu$ and $\delta_\omega \equiv \mathcal{M}$, this is equivalent to $\Upsilon(\frac{c_\theta}{C}) = \frac{\mu}{\mathcal{M}}\Psi(\frac{c_\theta/A_\theta}{\bar{N}})$. Substituting (B.2) into (B.1) shows that this also coincides with the planner's first-order condition pinning down relative firm sizes (B.1). Thus, conditional on C, the planner and the market choose the same relative firm-level allocations across consumption good producers.

Next, I use (B.2) to define the "entry" condition of the planner:

$$\int_{\theta^*}^{\infty} (\Upsilon_\theta(\frac{c_\theta}{C}) - \Psi_\theta(\frac{n_\theta}{N}) - \Psi_o(f_o/(L\bar{N})))dG(\theta) = \Psi_e(f_e/\,(LN)), \tag{B.6}$$

The free entry condition of the market, in turn, can be written:

$$\int_{\theta^*}^{\infty} \left(L \left(\mathcal{P}\Upsilon'_\theta(\frac{c_\theta}{C}) - \mathcal{W}\frac{1}{A_\theta}\Psi'_\theta(\frac{c_\theta}{\bar{N}A_\theta}) \right) c_\theta - \mathcal{W}\Psi'_o(\frac{f_o}{L\bar{N}})f_o \right) dG(\theta) = \mathcal{W}\Psi'_e(\frac{f_e}{L\bar{N}})f_e. \tag{B.7}$$

Under constant markups and markdowns, (B.6) and (B.7) coincide. To see this, divide (B.7) by L, and note, again, that when $\epsilon_\theta \equiv \mu$ and $\delta_{\omega'} \equiv \mathcal{M}$, then $\mathcal{P} = \frac{1}{C}\mu$ and $\mathcal{W} = \frac{1}{N}\mathcal{M}$. As a result, $\left(\mathcal{P}\Upsilon'_\theta(\frac{c_\theta}{C}) - \mathcal{W}\frac{1}{A_\theta}\Psi'_\theta(\frac{c_\theta}{NA_\theta})\right) c_\theta = \Upsilon(\frac{c_\theta}{C}) - \Psi(\frac{n_\theta}{N})$, $\mathcal{W}\Psi'_o(f_o/(LN))f_o/L = \Psi(\frac{f_o}{NL})$ and $\mathcal{W}\Psi'_e(f_e/(LN))f_e/L = \Psi(\frac{f_e}{NL})$. Thus, (B.6) and (B.7) provide the same restriction on entry. Analogous derivations imply that the planner's FOC pinning down θ^* (B.4) is equivalent to the market's selection equation in (25).

To establish the if part of the theorem note that free entry ensures that the planner and the market choose the same C. When firm-level allocations, the selection cutoff, and C coincide, the planner also chooses the same mass of entrants as the market.[32] By the previous arguments, this establishes that the planner and the market allocation coincide.

[32] This follows from noting that the free entry condition ensures $\chi(C, c_\theta, N, \theta^*) \equiv \Psi_e(\frac{f_e}{NL}) + \int_{\theta^*}^{\infty}(\Psi\left(\frac{c_\theta/A_\theta}{N}\right) + \Psi_o\left(\frac{f_o}{NL}\right))dG(\theta) = \int_{\theta^*}^{\infty}\Upsilon(\frac{c_\theta}{C})dG(\theta)$, so M_e adjusts so that $1 = M_e\chi(C, c_\theta, N, \theta^*)$. $U_C C = -\lambda_c M \int_{\theta^*}^{\infty}\Upsilon'(\frac{c_\theta}{C})\frac{c_\theta}{C}dG(\theta)$ is satisfied through adjustment of the multiplier so that $U_C C\mu = -\lambda_c$.

To prove the only if part, it is sufficient to show that conditional on C and, the planner and the market would always choose different output allocations to final good firms whenever $\exists\theta,$ s.t. $\epsilon_\theta \neq \mu_\theta$ or $\exists\omega'$ s.t. $\delta_{\omega'} \neq \mathcal{M}_{\omega'}$. Note that, in general, $\mathcal{P} = \frac{1}{C\mathbb{E}_{pc\epsilon}\left[\frac{1}{\delta_\theta}\right]}$ and $\mathcal{W} = \frac{1}{N\mathbb{E}_{wn\delta}\left[\frac{1}{\delta}\right]}$. Thus the per-capita demands in the market are generally given by: $\frac{1}{\mathbb{E}_{pc\epsilon}\left[\frac{1}{\epsilon_\theta}\right]}\Upsilon'(\frac{c_\theta}{C})\frac{1}{C}Y = p_\theta$, and $\frac{1}{\mathbb{E}_{wn\delta}\left[\frac{1}{\delta}\right]}\Psi'(\frac{n_\theta}{N})\frac{1}{N}Y = w_\theta$. Firm-level profit-maximization then can be written as

$$\Upsilon(\frac{c_\theta}{C}) = \frac{\mu_\theta}{\mathcal{M}_\theta}\Psi(\frac{c_\theta/A_\theta}{N})\frac{\mathbb{E}_{pc\epsilon}\left[\frac{1}{\epsilon_\theta}\right]\epsilon_\theta}{\mathbb{E}_{wn\delta}\left[\frac{1}{\delta}\right]\delta_\theta}. \tag{B.8}$$

Comparing equations (B.1) and (B.8), it is evident that a necessary condition for the market and planner allocations to coincide is that $\frac{\mu_\theta}{\mathcal{M}_\theta}\frac{\mathbb{E}_{pc\epsilon}\left[\frac{1}{\epsilon_\theta}\right]}{\mathbb{E}_{wn\delta}\left[\frac{1}{\delta}\right]} = 1$. □

Proof of Lemma 1

Proof. A reallocation of labor from $(\theta', \theta' + d\theta')$ to $(\theta, \theta + d\theta')$ that keeps overall labor supply N unchanged implies that for $d\log n_{\theta'} < 0$, the complementary increase in n_θ satisfies: $d\log n_\theta = -\frac{g(\theta')}{g(\theta)}\frac{w_{\theta'}n_{\theta'}}{w_{\theta'}n_{\theta'}}d\log n_{\theta'}$. Since $\frac{w_{\theta'}n_{\theta'}}{w_\theta n_\theta} = \frac{p_{\theta'}c_{\theta'}\frac{\mathcal{M}_{\theta'}}{\mu_{\theta'}}}{p_\theta c_\theta\frac{\mathcal{M}_\theta}{\mu_\theta}}$, the associated gain in the consumption utility index is given by

$$g(\theta')p_{\theta'}c_{\theta'}d\log n_{\theta'}d\theta' + g(\theta)p_\theta c_\theta d\log n_\theta d\theta' = -(\frac{\frac{\mathcal{M}_{\theta'}}{\mu_{\theta'}}}{\frac{\mathcal{M}_\theta}{\mu_\theta}} - 1)g(\theta')d\theta' d\log n_{\theta'}.$$

This is positive if, and only if, $\frac{\mathcal{M}_{\theta'}}{\mu_{\theta'}} > \frac{\mathcal{M}_\theta}{\mu_\theta}$,or equivalently, $\frac{\mu_\theta}{\mathcal{M}_\theta} > \frac{\mu_{\theta'}}{\mathcal{M}_{\theta'}}$. □

Proof of Lemma 2

Proof. Suppose we reduce output equally across all consumption varieties $d\log c_\theta = d\log\tilde{c} < 0$. The change in per-capita quantity consumed from the reallocation, provided that the change in welfare of of a reallocation that keeps selection unchanged, is equal to: $d\log C = \bar{\epsilon}d\log M + \mathbb{E}_{pc}d\log c_\theta$. To keep the labor supply index fixed, we require that $0 = \bar{\delta}d\log M + \mathbb{E}_{pc}\frac{\mathcal{M}_\theta}{\mu_\theta}d\log c_\theta$. This implies that $d\log M = -\frac{1}{\bar{\delta}}\mathbb{E}_{pc}\frac{\mathcal{M}_\theta}{\mu_\theta}d\log c_\theta$, and so we have that $d\log C = \mathbb{E}_{pc}(\frac{\bar{\epsilon}}{\bar{\delta}}\frac{\mathcal{M}_\theta}{\mu_\theta} - 1)(-d\log c_\theta)$. Since, $-d\log c_\theta = -d\log\tilde{c} > 0$, $d\log C > 0$ if, and only if, $\mathbb{E}_{pc}(\frac{\bar{\epsilon}}{\bar{\delta}}\frac{\mathcal{M}_\theta}{\mu_\theta} - 1) > 0$, which implies the condition stated in the main text. □

Proof of Lemma 3

Proof. Suppose the selection cutoff increases by $d\theta^* > 0$. As we keep relative employment unchanged across final, overhead and entry goods firms, the associated change in the mass of firms equals: $d\log M = \left(\frac{\mathcal{M}_{\theta^*}}{\mu_{\theta^*}}\frac{\delta_{\theta^*}}{\bar{\delta}} + (1 - \frac{\mathcal{M}_{\theta^*}}{\mu_{\theta^*}})\frac{\delta_o}{\bar{\delta}}\right)\omega^{pc}_{\theta^*}\frac{g(\theta^*)}{1-G(\theta^*)}d\theta^*$. The associated change in consumption utility, given that all pre-existing firm remain of the same size, is given by

$$d\log C = \left[(1 - \frac{\mathcal{M}_{\theta^*}}{\mu_{\theta^*}})(\frac{\delta_o}{\bar{\delta}\delta_{\theta^*}} - \frac{\bar{\epsilon}}{\bar{\delta}}) + \frac{\bar{\epsilon}}{\bar{\delta}} - \frac{\epsilon_{\theta^*}}{\delta_{\theta^*}}\right]\omega^{pc}_{\theta^*}\frac{g(\theta^*)}{1 - G(\theta^*)}d\theta^*.$$

With some manipulation it can be shown that tougher selection, $d\theta^* > 0$, increases welfare if, and only if,

$$\frac{\bar{\epsilon} - \epsilon_{\theta^*}}{\bar{\epsilon}} + \frac{(\delta_{\theta^*} - \bar{\delta})}{\bar{\delta}}\frac{\mathcal{M}_{\theta^*}}{\mu_{\theta^*}} + \frac{\delta_o - \bar{\delta}}{\bar{\delta}}(1 - \frac{\mathcal{M}_{\theta^*}}{\mu_{\theta^*}}) > 0,$$

which is equivalent to

$$\bar{\boldsymbol{\epsilon}} - \boldsymbol{\epsilon}_{\theta^*} + \frac{(1 - \boldsymbol{\mu}^e_{\theta^*})}{\bar{\delta}}(\frac{\delta_o}{\delta_{\theta^*}} - \bar{\epsilon}) > 0.$$

□

Proof of Proposition 1

Proof. Since $U(C^{\text{opt}}, N^{\text{opt}}) \geq U(C^{\text{mkt}}, N^{\text{mkt}})$, and $-\frac{U_C(C^{\text{opt}},N^{\text{opt}})}{U_N(C^{\text{opt}},N^{\text{opt}})}\frac{C^{\text{opt}}}{N^{\text{opt}}} < -\frac{U_C(C^{\text{mkt}},N^{\text{mkt}})}{U_N(C^{\text{mkt}},N^{\text{mkt}})}\frac{C^{\text{mkt}}}{N^{\text{mkt}}}$, the decentralized equilibrium is inefficient. To show the second part, note that aggregate labor supply of households is increasing in real wages under the stated condition: $\frac{\partial N}{\partial \frac{W^I}{P^I}} \propto U_C + N(U_{NC} + \frac{W^I}{P^I}(U_{CN} + U_{CC}) \propto 1 + \frac{U_{NC}N}{u_C} + \frac{CU_{CC}}{U_C}$. The assertion follows from the fact that inducing the optimal labor supply would require raising the real wage, $\frac{C^{\text{mkt}}}{N^{\text{mkt}}}$, by a factor $\frac{\bar{\epsilon}^{\text{opt}}}{\bar{\delta}^{\text{opt}}} > 1$. □

Proof of Proposition 2

Proof. The proof closely follows the arguments in Baqaee *et al.* (2022). That is, index equilibria (N, χ) by τ. When $\tau^* = 0$, we have that

$$\begin{aligned} \mathcal{L} = \log\frac{\mathcal{U}(\tau^*)}{\mathcal{U}(d\tau^*)} &\approx \frac{d\log\mathcal{U}}{d\tau}\Big|_{\tau=\tau^*} d\tau^* + \frac{1}{2}\frac{d^2\log U}{dt^2}\Big|_{\tau=\tau^*} (d\tau^*)^2 \\ &= \frac{1}{2}\frac{d^2\log U}{dt^2}\Big|_{\tau=\tau^*} (d\tau^*)^2 \\ &\approx \frac{1}{2}\frac{d\log\mathcal{U}}{d\tau}\Big|_{\tau=d\tau^*} d\tau^*. \end{aligned} \tag{B.9}$$

The second line uses the fact that, by the Envelope Theorem, the first derivative of

$\mathcal{U}$ with respect to the efficiency-inducing policy equals zero at the efficient allocation. The third line uses a first-order expansion. To calculate $\frac{d\log\mathcal{U}}{d\tau}$, one can use the fact that the distance to the frontier is given by integrating changes in welfare from the decentralized equilibrium at τ to the efficient allocation:

$$\log\frac{\mathcal{U}(\tau^*)}{\mathcal{U}(\tau)} = \int_{\tau}^{\tau^*}\left(\frac{\partial\log\mathcal{U}}{\partial\log N}\frac{\partial\log N}{\partial\nu} + \frac{\partial\log\mathcal{U}}{\partial\mathcal{X}}\frac{\partial\mathcal{X}}{\partial\nu}\right)d\nu. \tag{B.10}$$

Taking the derivative with respect to τ and applying the Envelope theorem,

$$\mathcal{L} \approx \log\frac{\mathcal{U}(\tau^*)}{\mathcal{U}(\tau)} = \frac{\partial\log\mathcal{U}}{\partial\log N}d\log\tau^N + \frac{\partial\log\mathcal{U}}{\partial\mathcal{X}}\frac{\partial\mathcal{X}}{\partial\ln N}d\log\tau^N + \frac{\partial\log\mathcal{U}}{\partial\mathcal{X}}d\tau^{\mathcal{X}}.$$

□

Proof of Proposition 3

Proof. First, I write the equilibrium conditions with changes in taxes. GDP per-capita continues to serve as the numeraire. The labor share, given a set of taxes on varieties τ^c_θ and jobs τ^n_ω is given by.

$$\Lambda_L = M\left\{\frac{1}{\tau^n_e}p_e\frac{f_e}{L} + (1-G(\theta^*)\frac{1}{\tau^n_o}p_o\frac{f_o}{L} + \int_{\theta>\theta^*} s_\theta\frac{\mathcal{M}_\theta}{\mu_\theta}\frac{1}{\tau^n_\theta\tau^c_\theta}\frac{dG(\theta)}{1-G(\theta^*)}\right\}.$$

The equilibrium conditions can be written:

$$(1-G(\theta^*)M\int_{\theta>\theta^*}\Upsilon_\theta(\frac{c_\theta}{C})\frac{dG(\theta)}{1-G(\theta^*)} = 1$$

$$M\left[\Psi_e(\frac{f_e}{NL}) + (1-G(\theta^*))\Psi_o(\frac{f_e}{NL}) + \int_{\theta>\theta^*}\Psi_\theta(\frac{n_\theta}{N})dG(\theta)\right] = 1$$

$$\tau^c_\theta\frac{\mu_\theta}{\mathcal{M}_\theta}\frac{w_\theta}{A_\theta}\Lambda_L = \mathcal{P}\Upsilon'(\frac{c_\theta}{C})$$

$$\tau^n_\theta\frac{\mathcal{M}_\theta}{\mu_\theta}p_\theta A_\theta\Lambda_L = \mathcal{W}\Psi'(\frac{n_\theta}{N})$$

$$\frac{1}{\mathcal{P}} = CM\int_{\theta>\theta^*}\frac{c_\theta}{C}\Upsilon'_\theta(\frac{c_\theta}{C})dG(\theta)$$

$$\frac{1}{\mathcal{W}} = NM\left[\frac{f_e}{NL}\Psi'_e(\frac{f_e}{NL}) + (1-G(\theta^*))\frac{f_o}{NL}\Psi'_o(\frac{f_o}{NL}) + \int_{\theta>\theta^*}\frac{n_\theta}{N}\Psi'_\theta(\frac{n_\theta}{N})dG(\theta)\right].$$

$$(1-\frac{\mathcal{M}_{\theta^*}}{\mu_{\theta^*}})\frac{s_{\theta^*}}{\tau^n_{\theta^*}\tau^c_{\theta^*}} = \frac{(1-G(\theta^*))M\Lambda_L p_o f_o}{L}$$

$$\tau_e^n p_e \Lambda_L = \mathcal{W}\Psi'(\frac{f_e}{NL}),\ \tau_o^n p_o \Lambda_L = \mathcal{W}\Psi'(\frac{f_e}{NL})$$

In changes,

$$d\log\frac{c_\theta}{C} = -\frac{\sigma_\theta\beta_\theta}{\sigma_\theta+\beta_\theta}\left[d\log(\tau_\theta^c\tau_\theta^n\frac{\mu_\theta}{\mathcal{M}_\theta}\Lambda_L\frac{\mathcal{P}}{\mathcal{W}}) + \frac{1}{\beta_\theta}d\log\frac{C}{N}\right]$$

$$d\log\frac{\mathcal{W}}{\mathcal{P}} = \mathbb{E}_s\left[\frac{1}{\sigma_\theta}\right]d\log\frac{C}{N} + \mathbb{E}_s\left[1-\frac{\mathcal{M}_\theta}{\mu_\theta}\right]\frac{1}{\mathcal{M}_f}d\log N$$

$$d\theta^* = -\gamma_{\theta^*}d\log(\tau_{\theta^*}^n\tau_{\theta^*}^c M\Lambda_L p_o)$$

$$d\log p_{\{e,o\}} = d\log\mathcal{W} - \frac{1}{\beta_{\{e,o\}}}d\log N - d\log\tau_{\{e,o\}}^n$$

$$\begin{aligned} d\log\tfrac{C}{N} = -\quad & s_{\theta^*}\left(\tfrac{\epsilon_{\theta^*}-\bar\epsilon}{\bar\delta} - \tfrac{\mathcal{M}_{\theta^*}}{\mu_{\theta^*}}\tfrac{\delta_{\theta^*}-\bar\delta}{\bar\epsilon} - (1-\tfrac{\mathcal{M}_\theta}{\mu_{\theta^*}})\tfrac{\delta_o-\bar\delta}{\bar\epsilon}\right)\tfrac{g(\theta^*)}{1-G(\theta^*)}d\theta^* \\ & +\mathbb{E}_s\left[(1-\tfrac{\bar\epsilon}{\bar\delta}\tfrac{\mathcal{M}_\theta}{\mu_\theta})d\log\tfrac{c_\theta}{C}\right]. \end{aligned}$$

Specializing these equations towards the efficient point and ignoring terms of order t^3,

$$d\log\tau^c\tau^n \approx -\log\frac{\varepsilon_{\theta^*}}{\delta_{\theta^8}}$$

$$d\log\tau^c\tau^n\frac{\mu_\theta}{\mathcal{M}_\theta} \approx -\log\frac{\mu_\theta}{\mathcal{M}_\theta}$$

$$d\log\tau_f^n \approx \log\delta_f$$

$$d\log\Lambda_L \approx -\mathbb{E}_s\left[1-\frac{\mathcal{M}_\theta}{\mu_\theta}\right]\sum_{f\in\{o,e\}}\omega_f\log\delta_f - \mathbb{E}_s\left[\frac{\mathcal{M}_\theta}{\mu_\theta}\log\frac{\epsilon_\theta}{\delta_\theta}\right]$$

$$\mathbb{E}_s\left[\left(1-\tfrac{\bar\epsilon}{\bar\delta}\tfrac{\mathcal{M}_\theta}{\mu_\theta}\right)d\log\tfrac{c_\theta}{C}\right] \approx \mathbb{E}_s\quad\left[\tfrac{\sigma_\theta\beta_\theta}{\sigma_\theta+\beta_\theta}\left(1-\tfrac{\bar\epsilon}{\bar\delta}\tfrac{\mathcal{M}_\theta}{\mu_\theta}\right)\left[\log\tfrac{\mathcal{M}_\theta}{\mu_\theta} - \mathbb{E}_s\left[(1-\tfrac{\mathcal{M}_\theta}{\mu_\theta})\log\tfrac{\bar\epsilon}{\bar\delta}\right]\right]\right]$$

$$d\theta^* = -\gamma_{\theta^*}\left[\frac{\varepsilon_{\theta^*}}{\delta_{\theta^*}} + \log\delta_o - \log\frac{\bar\epsilon}{\bar\delta}\right]$$

The result follows from imposing that $\log x \approx x - 1$. □

Proof of Theorem 2

Throughout, I make use of the following fact: Wage-bill weighted averages over outcomes of final good producers are given by $E_{wn}[x_\theta] = \mathbb{E}_{pc}\left[\frac{\mathcal{M}_\theta}{\mu_\theta}x_\theta\right]$. This follows from observing that $w_\theta n_\theta = \frac{\mathcal{M}_\theta}{\mu_\theta}p_\theta c_\theta$, so that $\frac{w_\theta n_\theta g((\theta)}{\int_\omega w_{\omega'}n_{\omega'}d\omega'} = \frac{p_\theta c_\theta\frac{\mathcal{M}_\theta}{\mu_\theta}}{\int_{\theta*}p_\theta c_\theta dG(\theta)}$. The last equality follows from the fact that total earnings equal total consumption spending.

First, we provide first-order expansions of all equilibrium conditions.

Setting up the system of equations

Differentiating the consumtion and labor indices, we obtain:

$$\mathbb{E}_s\left[\epsilon_\theta\right] d\log M - s_{\theta^*}\epsilon_{\theta^*}\frac{g(\theta^*)}{1-G(\theta^*)}d\theta^* + \mathbb{E}_s\left[d\log(\frac{c_\theta}{C})\right] = 0$$

$$\mathbb{E}_{wn}[\delta]d\log M + \mathbb{E}_s\left[\frac{\mathcal{M}_\theta}{\mu_\theta}d\log(\frac{n_\theta}{N})\right] - s_{\theta^*}\left(\frac{\mathcal{M}_{\theta^*}}{\mu_{\theta^*}}\delta_{\theta^*} + \left(1-\frac{\mathcal{M}_{\theta^*}}{\mu_{\theta^*}}\right)\delta_o\right)\frac{g(\theta^*)}{1-G(\theta^*)}d\theta^* - \mathbb{E}_s\left[1-\frac{\mathcal{M}_\theta}{\mu_\theta}\right](d\log NL) = 0$$

Differentating the wage and price aggregates:

$$-d\log\mathcal{P} = d\log C + d\log M - s_{\theta^*}\frac{g(\theta^*)}{1-G(\theta^*)}d\theta^* + \mathbb{E}_{wn}\left[\left(1-\frac{1}{\sigma_\theta}\right)d\log\left(\frac{c_\theta}{C}\right)\right]$$

$$-d\log\mathcal{W} = d\log N + \mathbb{E}_s\left[\left(1-\frac{1}{\sigma_\theta}\right)d\log\frac{n_\theta}{N}\right] - \mathbb{E}_s\left[1-\frac{\mathcal{M}_\theta}{\mu_\theta}\right]\frac{1}{\mathcal{M}_f}d\log NL - s_{\theta^*}\frac{g(\theta^*)}{1-G(\theta^*)}d\theta^*$$

where $\frac{1}{\mathcal{M}_f} \equiv \frac{(1-G(\theta^*)p_o f_o}{(1-G(\theta^*)p_o f_o + p_e f_e}\frac{1}{\mathcal{M}_o} + \frac{p_e f_e}{(1-G(\theta^*)p_o f_o + p_e f_e}\frac{1}{\mathcal{M}_e}$ is the average inverse markup in the entry and overhead goods sector. I also used the fact that the marginal firm θ^* makes no profits, so its cost incurred for variable labor equal excactly its payments for overhead.

Differentiating the inverse demand and supply functions facing firms:

$$d\log w_\theta - d\log\mathcal{W} = \frac{1}{\beta_\theta}d\log(\frac{n_\theta}{N})$$

$$d\log p_\theta - d\log\mathcal{P} = -\frac{1}{\sigma_\theta}d\log\left(\frac{c_\theta}{C}\right)$$

The relationship between prices and wages is given by:

$$d\log p_\theta - d\log w_\theta = d\log\mu_\theta - d\log\mathcal{M}_\theta$$

The production technology links per-capita output to per-capita employment:

$$d\log n_\theta = d\log c_\theta$$

Differentiating the markup and markdown equation, we obtain:

$$d\log\mathcal{M}_\theta = \frac{\gamma_\theta - 1}{\gamma_\theta}\frac{1}{\beta_\theta}d\log(\frac{n_\theta}{N})$$

$$d\log\mu_\theta = \frac{1}{\sigma_\theta}\frac{1-\rho_\theta}{\rho_\theta}d\log(\frac{c_\theta}{C})$$

Differentiating the free entry condition, we obtain:

$$\mathbb{E}_\pi\left[d\log\pi_\theta\right] + d\log L - \frac{\pi_{\theta^*}}{\mathbb{E}_s\left[1-\frac{\mathcal{M}_\theta}{\mu_\theta}\right]}\frac{g(\theta^*)}{1-G(\theta^*)}d\theta^* = d\log\mathcal{W} - \frac{1-\mathcal{M}_f}{\mathcal{M}_f}\left(d\log LN\right)$$

The total derivative of varibale profits is given by:

$$d\log\pi_\theta = d\log p_\theta + d\log\frac{c_\theta}{C} + d\log C + \frac{1}{\mu_\theta-\mathcal{M}_\theta}\left(d\log\mu_\theta - d\log\mathcal{M}_\theta\right)$$

Finally, differentiaing the selection cutoff condition:

$$d\log L + d\log\pi_{\theta^*} - \frac{1}{\zeta_{\theta^*}}d\theta^* = -\frac{1-\tilde{\mathcal{M}}_o}{\tilde{\mathcal{M}}_o}d\log(LN) + d\log\mathcal{W}$$

Solving the system

First, I express all equilibrium outcomes in terms of $d\log\mathcal{W}/\mathcal{P}$,$d\log\frac{C}{N}$,$d\log M$, $d\theta^*$.

Employment and production of firms I begin by deriving expressions for firm level quantities in terms of aggregate price and wage indices, as well as the consumption and labor supply indices:

$$d\log\frac{c_\theta}{C} = \sigma_\theta d\log\mathcal{P} - \sigma_\theta\left(\overbrace{\underbrace{\frac{1}{\sigma_\theta}\frac{1-\rho_\theta}{\rho_\theta}d\log(\frac{c_\theta}{C})}_{d\log\mu_\theta} - \underbrace{\frac{\gamma_\theta-1}{\gamma_\theta\beta_\theta}d\log\frac{n_\theta}{N}}_{d\log\mathcal{M}_\theta} + \underbrace{\frac{1}{\beta_\theta}d\log\frac{n_\theta}{N} + d\log\mathcal{W}}_{d\log w_\theta} - d\log A_\theta}^{d\log p_\theta}\right),$$

Using the fact that $d\log n_\theta = d\log c_\theta$, and $\frac{1}{\rho_\theta} + \frac{\sigma_\theta}{\beta_\theta}\frac{1}{\gamma_\theta} = \frac{\beta_\theta\gamma_\theta+\sigma_\theta\rho_\theta}{\rho_\theta\beta_\theta\gamma_\theta}$

$$d\log(\frac{c_\theta}{C}) = -\frac{\sigma_\theta\beta_\theta\rho_\theta\gamma_\theta}{\gamma_\theta\beta_\theta+\rho_\theta\sigma_\theta}d\log\mathcal{W}/\mathcal{P} - \frac{\sigma_\theta\rho_\theta}{\gamma_\theta\beta_\theta+\rho_\theta\sigma_\theta}d\log C/N \equiv -\chi_\theta d\log\mathcal{W}/\mathcal{P} - \frac{\chi_\theta}{\beta_\theta\gamma_\theta}d\log C/N \tag{B.11}$$

Changes in markups and markdowns, are given by:

$$\begin{aligned} d\log\tfrac{\mu_\theta}{\mathcal{M}_\theta} &= -\left(1-\rho_\theta\gamma_\theta\tfrac{\beta_\theta+\sigma_\theta}{\sigma_\theta\rho_\theta+\gamma_\theta\beta_\theta}\right)d\log\mathcal{W}/\mathcal{P} - \tfrac{1}{\beta_\theta}\left(1-\rho_\theta\tfrac{\beta_\theta+\sigma_\theta}{\sigma_\theta\rho_\theta+\gamma_\theta\beta_\theta}\right)d\log(\tfrac{C}{N}) \\ &\equiv (1-\Gamma_\theta)d\ln\boldsymbol{\mathcal{P}} - \tfrac{\gamma_\theta-\bar{\Gamma}_\theta}{\beta_\theta}d\ln\boldsymbol{\mathcal{A}} \end{aligned} \tag{B.12}$$

And quantity changes are given by, $\Sigma_\theta = \frac{\sigma_\theta+\beta_\theta}{\sigma_\theta+\beta_\theta}$

$$d\ln\tfrac{c_\theta}{C} = \quad -\Sigma_\theta\Gamma_\theta d\ln\tfrac{\mathcal{W}}{\mathcal{P}} - \tfrac{\Sigma_\theta\Gamma_\theta}{\beta_\theta\gamma_\theta}d\ln\tfrac{C}{N} \tag{B.13}$$

Relative price indices: The free entry condition implies,

$$\Lambda^{\mathcal{N}} d\log NL = d\log \mathcal{W}/\mathcal{P} - \Lambda^{\mathcal{Y}} d\log C/N, \tag{B.14}$$

where

$$\Lambda^{\mathcal{N}} \equiv \mathbb{E}_s\left[1-\frac{\mathcal{M}_\theta}{\mu_\theta}\right]\frac{1}{\mathcal{M}_f}, \quad \Lambda^{\mathcal{C}} \equiv \mathbb{E}_s\left[\frac{1}{\sigma_\theta}\right]$$

Free entry condition Substituting the expression for profits yields,

$$\begin{aligned}&\mathbb{E}_s\left[\left(1-\tfrac{\mathcal{M}_\theta}{\mu_\theta}\right)\tfrac{1}{\mathcal{M}_\theta}d\log\left(\tfrac{c_\theta}{C}\right)\right] + \mathbb{E}_s\left[\left(1-\tfrac{\mathcal{M}_\theta}{\mu_\theta}+\tfrac{1}{\mu_\theta}\right)d\log\tfrac{\mu_\theta}{\mathcal{M}_\theta}\right]\\ &= -\mathbb{E}_s\left[1-\tfrac{\mathcal{M}_\theta}{\mu_\theta}\right]\tfrac{1}{\mathcal{M}_f}d\log\left(NL\right) - \mathbb{E}_s\left[\left(1-\tfrac{\mathcal{M}_\theta}{\mu_\theta}\right)\tfrac{1}{\mathcal{M}_\theta}\right]d\log\tfrac{C}{N}\end{aligned} \tag{B.15}$$

Consumption and labor index:

$$\mathbb{E}_s\left[\epsilon_\theta\right]d\log M - s_{\theta^*}\epsilon_{\theta^*}\frac{g(\theta^*)}{1-G(\theta^*)}d\theta^* + \mathbb{E}_s\left[d\log(\frac{c_\theta}{C})\right] = 0 \tag{B.16}$$

$$\begin{aligned}\mathbb{E}_s\left[1-\tfrac{\mathcal{M}_\theta}{\mu}\right]d\log NL = \quad & \mathbb{E}_{wn}[\delta]d\log M + \mathbb{E}_s\left[\tfrac{\mathcal{M}_\theta}{\mu_\theta}d\log(\tfrac{c_\theta}{C})\right]\\ & +\mathbb{E}_s\left[\tfrac{\mathcal{M}_\theta}{\mu_\theta}\right]d\log\tfrac{C}{N} - s_{\theta^*}\left((1-\tfrac{\mathcal{M}_{\theta^*}}{\mu_{\theta^*}})\delta_o + \tfrac{\mathcal{M}_{\theta^*}}{\mu_{\theta^*}}\delta_{\theta^*}\right)d\theta^*\end{aligned}, \tag{B.17}$$

Proof of Theorem 2

Proof. I substitute $d\log\frac{c_\theta}{C}$ in all relevant equilibrium to solve for changes in $d\theta^*$, $d\log\frac{C}{N}$ and $d\log\frac{W}{P}$ as a function of the change in factor supply $d\log NL$. The change in selection equals:

$$d\theta^* = \iota_{\theta^*}\frac{\mu_{\theta^*}}{\mu_{\theta^*}-\mathcal{M}_{\theta^*}}d\log\frac{\mathcal{W}}{\mathcal{P}} - \iota_{\theta^*}\left(\frac{1}{\mathcal{M}_f}d\log NL - \frac{\mathbb{E}_s[\frac{1}{\sigma_\theta}]}{\mathbb{E}_s[1-\frac{\mathcal{M}_\theta}{\mu_\theta}]}d\log\frac{C}{N}\right)$$

Substituting for $d\log\frac{\mathcal{W}}{\mathcal{P}}$ using Equation (B.14) implies,

$$\tfrac{g(\theta)}{1-G(\theta^*)}d\theta^* = \quad -\iota_{\theta^*}(\tfrac{\boldsymbol{\mu}^e_{\theta^*}}{\boldsymbol{\mu}^e_{\theta^*}-1} - \mathbb{E}_s\left[\tfrac{\boldsymbol{\mu}^e_\theta-1}{\boldsymbol{\mu}^e_\theta}\right]^{-1})\left(\Lambda^{\mathcal{N}}d\log NL + \Lambda^{\mathcal{Y}}d\ln\tfrac{C}{N}\right)$$

Solving for $d\log M$ in (B.17), subtracing (B.17) from (B.16), and substituting the change in M gives

$$\begin{aligned}\mathbb{E}_s[\tfrac{\bar{\boldsymbol{\epsilon}}}{\boldsymbol{\mu}^e_\theta}]d\log\tfrac{C}{N} = \;& \bar{\boldsymbol{\epsilon}}\mathbb{E}_s\left[1-\boldsymbol{\mu}^e_\theta\right]d\log NL \\ & -s_{\theta^*}\left(\tfrac{\epsilon_{\theta^*}-\bar{\epsilon}}{\bar{\delta}}-\boldsymbol{\mu}^e_{\theta^*}\tfrac{\delta_{\theta^*}-\bar{\delta}}{\bar{\epsilon}}-(1-\boldsymbol{\mu}^e_{\theta^*})\tfrac{\delta_o-\bar{\delta}}{\bar{\epsilon}}\right)\tfrac{g(\theta^*)}{1-G(\theta^*)}d\theta^* \\ & +\mathbb{E}_s\left[(1-\tfrac{\bar{\boldsymbol{\epsilon}}}{\boldsymbol{\mu}^e_\theta})d\log\tfrac{c_\theta}{C}\right].\end{aligned}$$

Substituting for changes in firms' relative output $d\log\frac{c_\theta}{C}$ yields,

$$\begin{aligned}\mathbb{E}_s[\tfrac{\bar{\boldsymbol{\epsilon}}}{\boldsymbol{\mu}^e_\theta}]d\log\tfrac{C}{N} = \;& -\mathbb{E}_s\Big[(1-\tfrac{\bar{\boldsymbol{\epsilon}}}{\boldsymbol{\mu}^e_\theta})(\varSigma_\theta(\boldsymbol{\Lambda}^{\mathcal{Y}}+\tfrac{1}{\beta_\theta})-\varSigma_\theta\left(1-\Gamma_\theta\right)\boldsymbol{\Lambda}^{\mathcal{Y}}+\zeta^*_{|theta}\boldsymbol{\Lambda}^{\mathcal{Y}} \\ & \quad +\tfrac{\gamma_\theta-\Gamma_\theta}{\beta_\theta\gamma_\theta}+\iota_{\theta^*}\boldsymbol{\Lambda}^{\mathcal{Y}}\Big]d\log\tfrac{C}{N} \\ & +\left(\bar{\boldsymbol{\epsilon}}(1-\mathbb{E}_s\left[\tfrac{1}{\boldsymbol{\mu}^e_\theta}\right])-\mathbb{E}_s[(1-\tfrac{\bar{\boldsymbol{\epsilon}}}{\boldsymbol{\mu}^e_\theta})(\varSigma_\theta-\varSigma_\theta\left(1-\Gamma_\theta\right)+\zeta_{\theta^*}]\boldsymbol{\Lambda}^{\mathcal{N}}\right)d\log NL\end{aligned}$$

To rearrange this term to yield the statment in the theorem, I use the following identities.

$$\begin{aligned}&\mathbb{E}_s\left[\left(1-\bar{\boldsymbol{\epsilon}}(\tfrac{\mathcal{M}_\theta}{\mu_\theta})\right)\varSigma_\theta(\tfrac{1}{\mathcal{M}_\theta}-\mathbb{E}\left[\tfrac{1}{\mu_\theta}\right])\right] \\ &\qquad =-\left(\bar{\boldsymbol{\epsilon}}-1\right)Cov_s\left[\varSigma_\theta\tfrac{\mathcal{M}}{\mu},\tfrac{1}{\mu_\theta}\right]+\text{Cov}\left[\mathcal{M}_\theta,\tfrac{1}{\mu_\theta}\right] \\ &\qquad\quad +\mathbb{E}_s\left(1-\tfrac{\bar{\boldsymbol{\epsilon}}}{\boldsymbol{\mu_\theta}}\right)\mathbb{E}_s\left[1-\mathcal{M}_\theta\right] \\ &\mathbb{E}_s\left[\left(1-\tfrac{\bar{\boldsymbol{\epsilon}}}{\boldsymbol{\mu_\theta}}\right)\boldsymbol{\varSigma}_\theta\Lambda^{\mathcal{N}}\right]=-\left(\bar{\boldsymbol{\epsilon}}-1\right)Cov_s\left[\varSigma_\theta,\tfrac{\mathcal{M}_\theta}{\mu_\theta}\right] \\ &\qquad +\bar{\boldsymbol{\epsilon}}\text{Cov}\left[\mathcal{M}_\theta,\tfrac{\mathcal{M}_\theta}{\mu_\theta}\right]\tfrac{1}{\mathcal{M}_F}+\mathbb{E}_s\left(1-\tfrac{\bar{\boldsymbol{\epsilon}}}{\boldsymbol{\mu_\theta}}\right)\tfrac{\mathcal{M}_\theta}{\mathcal{M}_F} \\ &\qquad =-\left(\bar{\boldsymbol{\epsilon}}-1\right)Cov_s\left[\varSigma_\theta,\tfrac{\mathcal{M}_\theta}{\mu_\theta}\right]-\mathbb{E}_s\left(1-\tfrac{\bar{\boldsymbol{\epsilon}}}{\boldsymbol{\mu_\theta}}\right)\mathbb{E}_s\left[\mathcal{M}_\theta\right]\end{aligned} \tag{B.18}$$

Also, note that $1=\text{Cov}_s\left[\varSigma_\theta,\frac{1}{\mathcal{M}_\theta}\right]-\text{Cov}_s\left[\varSigma,\frac{1}{\mu_\theta}\right]-\mathbb{E}_s\left[\varSigma_\theta\right]\mathbb{E}_s\left[\frac{1}{\mathcal{M}_\theta}-\frac{1}{\mu_\theta}\right]$ □

Proof of Proposition 4

Proof. Changes in output are given by,

$$d\log C=\underbrace{d\log\mathcal{A}}_{\text{productivity}}+\underbrace{d\log N}_{\text{employment}}=\left(\frac{d\log\mathcal{A}}{d\log L}+1\right)d\log N$$

Totally differentiating the labor-leisure optimality condition of the household yields,

$$\left(-\varepsilon^{U_C}_C-\varepsilon^{U_N}_C-\varepsilon^{U_N}_N-\varepsilon^{U_C}_N\right)d\log N=\left(1+\varepsilon^{U_C}_C+\varepsilon^{U_N}_C\right)d\log\mathcal{A}+\frac{1}{1+\tau}d(1+\tau)$$

Changes in welfare

$$d\ln\mathcal{U}=\varepsilon^{U_C}_C\frac{d\ln\mathcal{A}}{d\ln N}d\ln N$$

Specializing these expressions to either KPR or GHH preferences, evaluating at the initial equilibrium $\tau = 0$,$d(1+\tau) = \frac{\epsilon}{\bar{\delta}} - 1$, and calculating in consumption equivalent terms, the welfare gain from correcting factor supply distortions thus equals,

$$\frac{1}{2}\frac{\partial \ln C}{\partial \ln U}\frac{\partial \ln \mathcal{U}}{\partial lnN} d \ln \tau^N = \begin{cases} \frac{1}{2} \cdot \frac{\varphi}{1+\varphi}\varepsilon_L^{\mathcal{A}}\left(\frac{\bar{\epsilon}}{\bar{\delta}} - 1\right) & \text{if } U(C,N) = \log C - \varphi\frac{N^{1+1/\varphi}}{1+1/\varphi} \\ \frac{1}{2} \cdot \varphi\frac{\varepsilon_L^{\mathcal{A}}}{1-\varphi\varepsilon_L^{\mathcal{A}}}\left(\frac{\bar{\epsilon}}{\bar{\delta}} - 1\right) & \text{if } U(C,N) = \log\left(C - \varphi\frac{N^{1+1/\varphi}}{1+1/\varphi}\right) \end{cases}$$

□

Existence and Uniqueness

C Calibration

C.1 Pass-through Identification

C.2 Details on the Calibration Implementation

Dolfen (2020) and Yeh *et al.* (2022) separately estimate markdowns and markups across German establishments using cost-minimization and production function approach to measuring market power following Hall (1988) and Loecker & Warzynski (2012).

I combine information within-sector sales distributions in Germany from Trottner (2020) with moments of firm-level markup and markdown estimates reported in Dolfen (2020) to obtain sales-weighted averages of markups and markdowns across final good producers. To do so, I use the fact that the functional form restrictions discussed in the main text imply that markups and markdowns are strictly increasing in firm sales. This allows matching the reported moments of the estimated markdown and markup distribution (25th, 50th, 75th, and 90th decile) to the establishment sales distribution without having to obtain direct access to the micro-level estimates.

A natural caveat of the calibration approach - theoretically and empirically - is that it imposes that markdowns and markups are, respectively, monotonous in firm sales. This is consistent with existing empirical work standard practices in the quantitative literature on markups, which typically imposes functional forms that imply markups are monotonous in relative firm size. I leave it to future work to directly estimate the shape of product demand and labor supply curves using micro-level data on wages, sales, and employment. It is also worth noting, again, that the theoretical results presented in this paper impose no functional form restrictions on demand and labor supply beyond the general set of assumptions laid out in Section 2.

The data used to calibrate the model is displayed in table C.1

Table C.1 Markdown and Markup Estimates for Germany

Share of Plants: θ	Cumulative Sales Share	Markdown $\mathcal{M}_\theta$	Markup μ_θ
0.21	0.005	0.92	1.052
0.37	0.011	0.92	1.053
0.48	0.025	0.912	1.043
0.58	0.033	0.904	1.054
0.76	0.09	0.88	1.062
0.83	0.14	0.87	1.07
0.89	0.31	0.85	1.08
0.95	0.48	0.82	1.11
0.98	0.76	0.75	1.18
0.999	0.982	0.71	1.24

To construct markups and markdowns across firm types $\theta \in [0,1]$, I construct the sales distribution by fitting a flexible spine function to:

$$S_\theta = \frac{\int_0^\theta s_\theta dG(\theta)}{d\theta}.$$

Using the fitted curve $\hat{S}_\theta$, I compute the sales density from,

$$\hat{s}_\theta = \frac{d\hat{S}_\theta}{d\theta}.$$

Then, I fit a fit a flexible spine function to the following objects taken from the data:

$$md(\theta) = \frac{\int_0^\theta s_\theta \frac{\mathcal{M}_\theta}{\mu_\theta} \mathcal{M}_\theta dG(\theta)}{\int_0^\theta s_\theta \frac{\mathcal{M}_\theta}{\mu_\theta} dG(\theta)}, mu(\theta) = \frac{\int_0^\theta s_\theta \mathcal{M}_\theta dG(\theta)}{\int_0^\theta s_\theta \frac{\mathcal{M}_\theta}{\mu_\theta} dG(\theta)},$$

and recover markdowns and markups from,

$$\mathcal{M}_\theta = \frac{\int_0^\theta \hat{s}_{\theta'} d\theta'}{\hat{s}_\theta} \frac{d\widehat{md}(\theta)}{d\theta} + \widehat{md}(\theta), \mu_\theta = \frac{\int_0^\theta \hat{s}_{\theta'} d\theta'}{\hat{s}_\theta} \frac{d\widehat{mu}(\theta)}{d\theta} + \widehat{mu}(\theta).$$

Given the fitted values $\{\frac{d\mathcal{M}_\theta}{d\theta}, \mathcal{M}_\theta, \mu_\theta, \frac{d\mu_\theta}{d\theta}, s_\theta\}_\theta$,I solve for the pass-throughs $\{\gamma_\theta, \rho_\theta\}_\theta$ using the equation given in the main text. To solve for household rents, I solve the differential equations informing worker and firm rents using the Runge-Kutta method.

D Extensions

D.1 HSA labor supply and product demand

Here, I re-derive the main theoretical results for alternative labor supply and product demand systems. As the generalized Kimball product demand system used in the main text, the "homothetic with a single aggregator - H.S.A. - product demand system was introduced by Matsuyama & Ushchev (2022) as a homothetic generalization of the CES demand system. I show how the H.S.A. can be used to generate a labor supply system with variable wage elasticities.

D.1.1 Setup

Households There is a population of L identical households. Each household chooses the labor supply N and consumption $\mathcal{Y}$ to maximize utility given by,

$$\mathcal{U} = \mathcal{U}(\mathcal{Y}, \mathcal{N}).$$

Conditional on N, households choose how to allocate labor across jobs $\omega \in \Theta \cup \{o, e\}$, and how much to consume of each available consumption variety $\theta \in \Theta$. The share of a household's total earnings from each job ω is,

$$w_\omega n_\omega = l_\omega(\frac{w_\omega}{W}), \tag{D.1}$$

where n_ω denotes per-capita hours supplied to job ω at wage w_ω, and I is per-capita income. The earnings function $l_\theta(.)$ is increasing and satisfies $\lim_{x\to 0} l_\theta(x) = 0$ and $\lim_{x\to\infty} l_\theta(x) = \infty$.[33] The wage aggregator $\mathcal{W}$ is implicitly defined by the requirement that total earning shares sum to 1.

$$\int_\Omega l_\theta\left(\frac{w_\omega}{W}\right) dM^e(\omega) = 1,$$

where $M^e(\omega)$ denotes the measure of jobs of type ω. Denoting $\mathcal{W}^I$ the ideal price wage, nominal GDP is chosen as the numeraire., that is $I = \mathcal{W}^I N = 1$.

Households' expenditure share on a variety of type θ is given by,

$$\frac{p_\theta y_\theta}{I} = s_\theta(\frac{p_\theta}{\mathcal{P}}), \tag{D.2}$$

[33] It is straightforward to extend the arguments in Matsuyama & Ushchev (2017) to show that these conditions guarantee that the labor supply system can be rationalized by a monotone, concave, continuous, and homothetic rational preference relation.

where y_θ and p_θ denote the per-capita consumption and price of variety θ. $s_\theta(.)$ is a strictly decreasing function satisfying $\lim_{x\to 0} s_\theta(x) = \infty$ and $\lim_{x\to\infty} s_\theta(x) = 0$, and $\mathcal{W}$ is a price aggregator implictly defined by,

$$\int_\Theta s_\theta\left(\frac{p_\theta}{\mathcal{P}}\right) dM^C(\theta) = 1,$$

where $M^C(\theta)$ denotes the measure of varieties of type θ. Denote $\mathcal{P}^I$ the ideal consumption price index solving the expenditure minimization problem of the household.

Given P and W, households choose $\mathcal{Y}$ and N by setting,

$$-\frac{U_N}{U_\mathcal{Y}} = \frac{\mathcal{W}^I}{\mathcal{P}^I} = \frac{C}{N}.$$

Final Good Firms Equations (D.1) and (D.2) define per-capita labor supply and product demand to each firm. The labor supply elasticity faced by an employer offering wage w is given by

$$\beta_\omega(\frac{w}{W}) = \frac{\partial \log n_\omega}{\partial \log w_\omega} = \frac{\frac{w}{W} l'_\theta(\frac{w}{W})}{l_\theta(\frac{w}{W})} - 1, \tag{D.3}$$

while the price demand elasticity faced by firms of type θ is given by,

$$\sigma_\theta(\frac{p}{P}) = -\frac{\partial \log y_\theta}{\partial \log p_\theta} = 1 - \frac{\frac{p}{P} s'_\theta(\frac{p}{P})}{s_\theta(\frac{p}{P})}. \tag{D.4}$$

The market structure is the same as in the model presented in the main text, so prices and wages offered by a final good firms of type θ solve,

$$p_\theta = \frac{\mu_\theta(\frac{p}{P})}{\mathcal{M}_\theta(\frac{w}{W})} \frac{w_\theta\left(\frac{w}{W}\right)}{A_\theta}, \tag{D.5}$$

where $\mu_\theta(\frac{p}{P}) = \frac{\sigma_\theta(\frac{p}{P})}{\sigma_\theta(\frac{p}{P})-1}$ and $\mathcal{M}_\theta(\frac{w}{W}) = \frac{\beta_\theta(\frac{w}{W})}{\beta_\theta(\frac{w}{W})+1}$. Wages are pinned down by the condition that a firm's per-capita output y_θ satisfies household demand,

$$\frac{s_\theta(\frac{p_\theta}{P})}{p_\theta} = \frac{l_\theta(\frac{w_\theta}{W})}{w_\theta A_\theta}. \tag{D.6}$$

Assuming that firms can be ordered by type so that variable profits are increasing, the marginal entrant θ^* is determined by the exit condition:

$$Lp_{\theta^*} c_{\theta^*}\left(1 - \frac{\mathcal{M}_{\theta^*}}{\mu_{\theta^*}}\right) = p_o f_o. \tag{D.7}$$

The mass of entrants M is pinned down by the free entry condition,

$$\int_{\theta^*}^{\sup(\text{support}(G))} \left[Lp_\theta c_\theta \left(1 - \frac{\mathcal{M}_\theta}{\mu_\theta}\right) - p_o f_o \right] = p_e f_e, \tag{D.8}$$

and $dM^e(\theta) = dM^C(\theta) = M 1_{\theta > \theta^*} g(\theta) d\theta$.

Entry and overhead producers As in the main text, entry and overhead goods are indivisible and homogeneous, and supplied under perfect competition with free entry. Thus, the mass of overhead producing firms equals $dM^e(o) = M(1 - G(\theta^*))$, and the price of overhead goods solves,

$$\frac{f_o}{LN} = l_o(\frac{w_o}{W})/w_o. \tag{D.9}$$

Similarly, the mass of entry good producing firms equals $dM^E(e) = M$, and the price of the entry good solves,

$$\frac{f_e}{LN} = l_e(\frac{w_e}{W})/w_e. \tag{D.10}$$

Equilibrium Given $\{L, f_o, f_e, G(.)\}$ an equilibrium consists of labor allocations, a selection cutoff, a mass of entrants, and aggregate consumption and employment, $(\{n_\theta\}_{\theta \in \theta}, n_o, n_e, \theta^*, M, C, N)$, so that given prices $(\{p_\theta, w_\theta\}_{\theta \in \Theta}, p_o, p_e)$, housholds maximize utility taking prices as given, firms maximize profits taking the wage and price aggregate as given, and markets clear.

D.1.2 Concepts

As in the model used in the main text, worker and consumer surpluses, as well as wage and price passthroughs, can be expressed in terms of primitives of the model.

Household surplus The infra-marginal consumption surplus ϵ_θ is defined as the area under the demand curve to sales for variety θ,

$$\epsilon_\theta = \frac{\int_0^{c_\theta} p_\theta(y) dy}{p_\theta c_\theta} = 1 + \frac{\int_{p_\theta/\mathcal{P}}^{\infty} \frac{s_\theta(\xi)}{\xi} d\xi}{s_\theta(\frac{p_\theta}{\mathcal{P}})}.$$

The infra-marginal labor surplus δ_ω, in turn, is defined as 1 minus the area under the labor supply curve to earnings for job ω,

$$\delta_\omega = 1 - \frac{\int_o^{n_\theta} w_\omega(n) dn}{w_\omega n_\omega} = \frac{\int_0^{w_\omega/\mathcal{W}} \frac{l_\omega(\xi)}{\xi} d\xi}{l_\omega(\frac{w_\omega}{\mathcal{W}})}.$$

Naturally, $\delta_\omega \leq 1$, and $\epsilon_\theta \geq 1$,where the equalities hold whenever jobs or varieties are perfect substitutes.

Pass-throughs The pass-through of shocks to the marginal revenue product of labor into wages offered by ω,γ_ω, is given by,

$$\gamma_\omega(\frac{w}{W}) = \frac{\partial \log w_\omega}{\partial mrpl_\omega} = \frac{1}{1 - \frac{\frac{w}{W}\mathcal{M}'_\omega(\frac{w}{W})}{\mathcal{M}(\frac{w}{W})}},$$

while the pass-through of shocks to marginal cost into prices offered by type θ, ρ_θ, is given by,

$$\rho_\theta\left(\frac{p}{\mathcal{P}}\right) = \frac{\partial \log p_\theta}{\partial \log mc_\theta} = \frac{1}{1 - \frac{\frac{p}{\mathcal{P}}\mu'_\theta(\frac{p}{\mathcal{P}})}{\mu_\theta(\frac{p}{\mathcal{P}})}}.$$

D.1.3 Efficiency

The following theorem provides the analogue to Theorem 1 in the main text.

Theorem 3. *1. Suppose that preferences $\mathcal{U}$ are such that the Frisch elasticity equals 0. Then the market allocation is efficient, if, and only if, for all $\omega \in \Theta \cup \{o, e\}$ $l_\omega(x) = a_\omega x^{1+\beta}$ and for all $\theta \in \Theta$,$s_\theta(x) = b_\theta x^{1-\sigma}$, where $a_\omega, b_\theta \in \mathbb{R}_0^+$,$\sigma > 1, \beta > 1$. When the allocation is inefficient, the relevant sources of inefficiency remain characterized by lemmas 1, 2, and 3.*

2. If factor supply N is elastic, the market allocation is inefficient. If factor supply is strictly upward-sloping, aggregate labor supply in the market is strictly less than in the optimal allocation.

D.1.4 Propagation equations

The propagation equations for the model with HSA-type preferences are given by

$$d\log M - s_{\theta^*}\frac{g(\theta^*)}{1 - G(\theta^*)}d\theta^* - \mathbb{E}_s\left[\frac{1}{\mu_\theta}\sigma_\theta d\ln\frac{p_\theta}{P}\right] = 0,$$

$$d\log M - s_{\theta^*}\frac{g(\theta^*)}{1 - G(\theta^*)}d\theta^* + \mathbb{E}_s\left[\frac{\mathcal{M}_\theta}{\mu_\theta}(1 + \beta_\theta)d\log(\frac{w_\theta}{W})\right] - \mathbb{E}_s\left[1 - \frac{\mathcal{M}_\theta}{\mu_\theta}\right]\frac{1 + \beta_F}{\beta_F}d\ln\frac{NL}{W} = 0,$$

$$d\log\mu_\theta = \frac{\rho_\theta - 1}{\rho_\theta}d\log\frac{p_\theta}{\mathcal{P}}$$

$$d\log\mathcal{M}_\theta = \frac{\gamma_\theta - 1}{\gamma_\theta}d\log\frac{w_\theta}{\mathcal{W}}$$

$$-\sigma_\theta d\log\frac{p_\theta}{P} = \beta_\theta d\log\frac{w_\theta}{W} - d\log\frac{W}{P}$$

$$\beta_F d\log\frac{w_F}{W} = -d\log NL + d\ln W$$

$$d\log C = \bar{\epsilon} d\log M - s_{\theta^*}\epsilon_{\theta^*}\frac{g(\theta^*)}{1-G(\theta^*)}d\theta^* - \mathbb{E}_s\left[\sigma_\theta d\ln\frac{p_\theta}{P} + d\ln P\right]$$

$$-d\log N = d\ln W - d\textstyle\int_\Omega \left[\int_{w_\omega}^\infty \frac{l_\omega(\xi)}{\xi}d\xi\right] dM^E(\omega)$$

$$\begin{aligned} d\ln N =& \bar{\delta} d\ln M - s^w_{\theta^*}\delta^e_{\theta^*} + \mathbb{E}_s\left[\frac{\mathcal{M}_\theta}{\mu_\theta}\left(\beta_\theta d\ln\frac{w_\theta}{W} - d\ln W\right)\right] \\ &- \mathbb{E}_s\left[1 - \frac{\mathcal{M}_\theta}{\mu_\theta}\right] d\ln NL \end{aligned}$$

Defining$\Gamma^p_\theta \equiv \rho_\theta \frac{\beta_\theta\gamma_\theta+1}{\beta_\theta\gamma_\theta+\sigma_\theta\rho_\theta}, \Gamma^w_\theta \equiv \gamma_\theta\frac{(\sigma_\theta\rho_\theta-1)}{\beta_\theta\gamma_\theta+\sigma_\theta\rho_\theta}$ the system of equations reduces to:

Change in Prices

$$d\ln p_\theta = (1-\Gamma^p_\theta)d\ln P + \Gamma^p_\theta d\ln W$$

$$d\ln\frac{w_\theta}{W} = \gamma_\theta\frac{(1-\sigma_\theta\rho_\theta)}{\beta_\theta\gamma_\theta+\sigma_\theta\rho_\theta}d\ln\frac{W}{P} \Leftrightarrow d\ln w_\theta = \frac{\gamma_\theta(1+\gamma_\theta\beta_\theta)+\sigma_\theta\rho_\theta(1-\gamma_\theta)}{\beta_\theta\gamma_\theta+\sigma_\theta\rho_\theta}d\ln W + \gamma_\theta\frac{\sigma_\theta\rho_\theta - 1}{\beta_\theta\gamma_\theta+\sigma_\theta\rho_\theta}d\ln P$$

$$d\ln p_\theta - d\ln w_\theta = \left(1 - \rho_\theta\gamma_\theta\frac{\beta_\theta+\sigma_\theta}{\beta_\theta\gamma_\theta+\sigma_\theta\rho_\theta} + \frac{\gamma_\theta-\rho_\theta}{\beta_\theta\gamma_\theta+\sigma_\theta\rho_\theta}\right)d\ln\frac{P}{W}$$

Competition

$$\mathbb{E}_s\left[\frac{1}{\mu_\theta}\right]d\ln\frac{W}{P} = \mathbb{E}_s\left[1-\frac{\mathcal{M}_\theta}{\mu_\theta}\right]\frac{1}{\mathcal{M}_F}d\ln\frac{NL}{W}$$

Welfare

$$d\log C = \bar{\epsilon} d\log M - s_{\theta^*}\epsilon_{\theta^*}\frac{g(\theta^*)}{1-G(\theta^*)}d\theta^* - \mathbb{E}_s\left[\sigma_\theta d\ln\frac{p_\theta}{P} + d\ln P\right]$$

$$\begin{aligned} d\ln N =& \bar{\delta} d\ln M - s^w_{\theta^*}\delta^e_{\theta^*} - \mathbb{E}_s\left[\frac{\mathcal{M}_\theta}{\mu_\theta}\left(\sigma_\theta d\ln\frac{p_\theta}{P} + d\ln P\right)\right] \\ &- \mathbb{E}_s\left[1 - \frac{\mathcal{M}_\theta}{\mu_\theta}\right] d\ln NL \end{aligned}$$

$$
\begin{aligned}
d\ln\frac{C}{N} &= \left(\bar{\epsilon}-\bar{\delta}\right) d\ln M - s_{\theta^*}\left(\epsilon_{\theta^*}-\delta^e_{\theta^*}\right)\frac{g(\theta^*)}{1-G(\theta^*)}d\theta^* \\
&\quad - \mathbb{E}_s\left[\left(1-\frac{\mathcal{M}_\theta}{\mu_\theta}\right)\left(\sigma_\theta d\ln\frac{p_\theta}{P}+d\ln P\right)\right]+\mathbb{E}_s\left[1-\frac{\mathcal{M}_\theta}{\mu_\theta}\right]d\ln NL \\
&= \left(\bar{\epsilon}-\bar{\delta}\right) d\ln M - s_{\theta^*}\left(\epsilon_{\theta^*}-\delta^e_{\theta^*}\right)\frac{g(\theta^*)}{1-G(\theta^*)}d\theta^* \\
&\quad - \mathbb{E}_s\left[\left(1-{}^1\!/\boldsymbol{\mu}^e_\theta\right)\left(\sigma_\theta\Gamma^p_\theta\right)\right]d\ln\frac{W}{P}+\mathbb{E}_s\left[1-{}^1\!/\boldsymbol{\mu}^e_\theta\right]d\ln NL/P
\end{aligned}
$$

$$
d\log M = s_{\theta^*}\frac{g(\theta^*)}{1-G(\theta^*)}d\theta^* + \mathbb{E}_s\left[\frac{1}{\mu_\theta}(1-\sigma_\theta\Gamma^P_\theta)\right]d\ln\frac{W}{P} - \mathbb{E}_s\left[\frac{1}{\mu_\theta}\right]d\ln\frac{W}{P},
$$

$$
d\log M - s_{\theta^*}\frac{g(\theta^*)}{1-G(\theta^*)}d\theta^* - \mathbb{E}_s\left[\frac{1}{\mu_\theta}\left(\sigma_\theta\right)\right] - \mathbb{E}_s\left[1-\frac{\mathcal{M}_\theta}{\mu_\theta}\right]\frac{1+\beta_F}{\beta_F}d\ln\frac{NL}{W} = 0
$$

D.2 Variable Elasticity (VES) labor supply

The labor supply system used in the main text has two advantages: First, it is homothetic. Therefore, it has a natural microfoundation based on aggregating individual labor supply decisions of workers, and can easily be embedded into richer models. Second, it allows each firms' markdown and pass-through to vary as a function of a firms' size relative to its competitiors, while nesting constant markdowns and full pass-through across firms as a parametric special case. This section shows that an alternative labor supply system that delivers the latter but not the former advantage is that generated by variable elasticity of substitution preferences (as introduced by Dixit & Stiglitz (1977)). Let the labor disutility index N be given by:

$$
N = \int_{\omega\in\Omega}\Psi_\omega(n_\omega)d\omega.
$$

As before, the labor disutility indices $\Psi_{\omega'}(.)$ are strictly increasing and convex. Note that CES is, again, a special case of the above preferences for employment opportunities. In this case, the per-capita labor supply to employer ω' is given by:

$$
n_{\omega'} = \mathcal{S}_\omega\left(w_\omega\mathcal{W}\right),
$$

where $\mathcal{S}_\omega(.) \equiv (\Psi'_\omega)^{-1}(.)$ and $\mathcal{W} \equiv \frac{\int_{\Omega'}\Psi'_{\omega'}(n_{\omega'})n_{\omega'}d\omega'}{Y}$. $\mathcal{W}$ is a wage index that mediates monopsonistic competition among firms. Indeed, firms operating on different parts of the labor supply curve face different labor supply elasticities $\beta_\omega = \frac{\partial\log S_{\omega'}(w_\omega\mathcal{W})}{\partial\log(w_\omega\mathcal{W})}$ so long as the labor disutility indices are not CES.

For brevity, I assume that the consumption utility index C is given by a CES aggregator with elasticity of substitition σ. The market allocation can be characterized through the exact same set of equations that defined a decentralized equilibrium in the benchmark model described in Section 2.

The following result confirms that efficiency in a VES economy is tied to exactly the same conditions that characterized efficient allocations in the benchmark economy.

Proposition 5. *In an economy with inelastic aggregate, and firm-level VES labor supply and constant markups, the decentralized equilibrium is efficient if, and only if, $\Psi_{\omega'}(x) = b_{\omega'} x^{\frac{\beta+1}{\beta}}$, where $\beta > 1$, and $b_{\omega'} \in \mathbb{R}^{+}$.*

Unsurprisingly, all the intuitions underlying the main result characterizing efficiency in the benchmark model apply in the economy with VES labor supply, too. Specifically, private and social profit margins are still instrumental for characterizing efficient outcomes and understanding the nature of distortions. Only in the special case of isoelastic labor supply, private incentives are aligned with social incentives for production, and the appropriability and business stealing externalities exactly offset each other. When markdowns vary across firms, distortions in private and social incentives are vary across employers, and the distribution of these distortions characterize misallocation in allocations, entry, and exit. In fact, the same sufficient statistics discussed earlier characterize distortions and help sign the impact of industrial policy.

D.3 Heterogeneous Factors

Households The economy is populated by $s = 1, 2, ..., S$ worker groups. Each worker group consists of L_s households. Labor markets are segmented by worker group s.

To isolate the role of monopsony, I assume that households have CES preferences over consumption varieties with elasticity of substitution σ. Given prices and (group-specific) $w_{s,\omega'}$, hosueholds belonging to group s choose labor supply n^s_ω and consumption c^s_θ so as to maximize utility given by $U(C^s, \bar{N}^s)$, where

$$C_s = \left(\int_\theta (c_{s,\omega})^{(\sigma-1)/\sigma} d\omega \right)^{\sigma/(\sigma-1)}, \quad 1 = \int_\Omega \Psi^s_\omega\left(\frac{n_{s,\omega}}{\bar{N}^s}\right) dM^E(\omega),$$

where $\bar{N}_s$ denotes the fixed amount of labor supplied by group s. Note that the labor disutility index Ψ^s_ω now varies across worker groups s and employers ω. Inverse per-capita labor supply of group s, in turn, is given by:

$$\frac{w_{s,\omega'}}{\mathcal{W}_s} = \Psi^s_{\omega'}\left(\frac{n_{s,\omega'}}{\bar{N}_s}\right) Y_s,$$

where Y_s is the total earnings of worker groups s. The wage index is defined analogously to the model layed out in Section 2. Let $\beta_{s,\omega}$ denote the elasticity of labor supply of workers of type s to employer ω'.

Production Firms wishing to produce consumption goods purchase entry goods at price $p_e f_e$ to draw a type θ from a pdf $g(\theta)$ with cdf $G(\theta)$. After paying overhead costs of $p_o f_o$, firms produce output using a Cobb-Douglas production function given by:

$$y_\theta = A_\theta \prod_s n_{s,\theta}^{\alpha_s},$$

where $\sum_s \alpha_s = 1$.

Profit-maximization implies that offered wages to employees of type s apply a markdown to the marginal revenue product of labor:

$$w_{\theta,s} = \frac{\beta_{s,\theta}}{\beta_{s,\theta}+1} mrpl_{s,\theta} \equiv \mathcal{M}_{s,\theta} mrpl_{s,\theta}.$$

Note that markdowns now potentially vary across both worker types and firms. In other words firms may have different degrees of labor market power in each labor market. Prices apply a markup $\mu = \frac{\sigma}{\sigma-1}$ over marginal cost and are given by:

$$p_\theta = \frac{\mu}{\tilde{\mathcal{M}}_\theta} \prod_s w_{s,\theta}^{\alpha_s} / A_\theta,$$

and $\tilde{\mathcal{M}}_\theta = \prod_s (\mathcal{M}_{\theta,s})^{\alpha_s}$ is the firms' effective markdown. Firms net of overhead are given by $\pi_\theta = L(1 - \frac{\mu}{\tilde{\mathcal{M}}_\theta}) - p_o f_o$.

The zero profit condition pins down the cutoff for exit, $\pi_{\theta^*} = 0$, and the free entry condition is given by $p_e f_e = \int_{\theta^*}^{\infty} \pi_\theta dG(\theta)$.

Entry and overhead goods are produced under perfect competition by homogeneous firms endowed with the same Cobb-Douglas production technologies as final good firms. Again, firms in this sector price at marginal cost, but hire workers in the same labor market as final good firms.

Equilibrium A competitve equilibrium is defined analogously to the benchmark model by a mass of entrants, an exit cutoff, as well as allocations of workers across firms such that the free entry and zero profit conditions hold, firms maximize profits, households maximize utility, and markets clear.

Efficiency The social planner seeks to maximize a utilitarian welfare function that applies equal weights to the utility of every household in the economy.[34] The follow-

[34] For the detailed description of the planner's problem see See Appendix B.

ing result shows that the efficiency of the decentralized equilibrium is tied to homogeneous labor market power across both firms and labor markets.

Proposition 6. *In the economy with heterogeneous worker types and constant markups, the decentralized equilibrium is efficient if, and only if,* $\Psi^s_{\omega'}(x) = b_{s,\omega'} x^{\frac{\beta+1}{\beta}}$*, where* $\sigma \in (0,1)$*,* $\beta > 1$*, and* a_ω*,*$b_{\omega'} \in \mathbb{R}^+$*.*

Proposition 6 shows tthat efficiency in an economy with heterogeneous worker types and Cobb-Douglas production technologies requires that firms wage markdowns are the same in all labor segments. It is not sufficient for firms to have homogeneous degrees of labor market power within labor markets. Intuitively, differences in labor market power across markets distort firms' relative labor demands for different worker groups. To see the intuition more formally, recall that efficiency requires equalizing aggregate social and private profit margins. Private profit margins, determining incentives for entry/exit, are captured by $\mu/\tilde{\mathcal{M}}_\theta$. Social benefits, in turn, are captured by $\mu/\tilde{\delta}_\theta$, where $\tilde{\delta}_\theta \equiv \sum_s \alpha_s \epsilon_{s,\theta}$. Now, suppose firms have the same degree of labor market power within each labor market, but that market power might differ across markets so that $\mathcal{M}_{s,\theta} = \epsilon_{s,\theta}$. In this case, private and social profit margins are not aligned, given the simple and geometric average do not coincide: $\sum_s \alpha_s \mathcal{M}_{s,\theta} \neq \prod_s \mathcal{M}^{\alpha_s}_{s,\theta}$.

This result shows that in the model, heterogeneity in labor market power across *either* worker groups or firms results in misallocations. This highlights that measurement and quantification of misallocations caused through monopsony requires careful understanding and measurement of the nature and degree of labor market power both across and within labor markets.