Misallocations in Monopsonistic Labor Markets

Fabian Trottner*
UC, San Diego

February 2022

Preliminary (latest draft here) - Comments welcome

Abstract

When is monopsonistic competition between firms a source of aggregate misallocations? To study this question, I develop a heterogeneous firm model featuring endogenously variable markdowns and markups, entry, and exit. I show how monopsony power manifests in aggregate distortions depending on how rents in labor and product markets affect private and social production incentives. When markdowns are homogeneous, monopsony power is not a distortion, while social efficiency further requires homogeneous markups. In general, elasticities related to labor supply and product demand determine how distortions from monopsonistic competition manifest in labor allocations across firms, the number of entrants, and selection. The model, thereby, highlights that assessing misallocations from monopsony requires accounting for firms' product market power and vice versa. I use the model to show that these considerations have first-order implications for policy design and how market integration, i.e., trade, affects welfare and misallocations.

*Contact: ftrottner@ucsd.edu

1 Introduction

Empirical work finds that differences in labor market power within industries are pervasive,[1] raising concerns about the welfare consequences of imperfect competition in labor markets. This paper investigates how labor market power - the ability of firms to set wages - affects the efficiency of resource allocations across heterogeneous firms, focusing on three questions. First, when does monopsonistic competition between firms lead to distortions? Second, how does labor market power interact with product market power in shaping aggregate misallocations? Third, what are suitable policies to address potential misallocations?

To study these questions, I develop and embed a tractable model of monopsonistic competition with variable markdowns into a heterogeneous firm model featuring variable markups, entry, and exit. I show that efficiency in this environment requires homogeneous markdowns and markups, while micro-level heterogeneity in markdowns always results in distortions. Distortions manifest via relative firm sizes, entry, and exit can be characterized in terms of firm-level elasticities related to labor supply and product demand, providing new insights into the determinants of resource misallocation in an economy where firms exercise both labor and product market power. Finally, I leverage the model to trace out policy implications, showing how the allocational and welfare effects of taxation and market integration, i.e., trade, are shaped by monopsonistic competition and its interactions with firm product market power.

The labor market model generalizes benchmark frameworks of static monopsony described in, e.g., Manning (2011) and Card *et al.* (2018), providing a novel microfoundation for upward-sloping labor supply curves that feature variable wage elasticities even when firms are atomistic.[2] Labor market power arises as jobs are differentiated from the perspective of workers. As a result, wages depend on firm size, are lower than the competitive level, and markdowns endogenously vary across employers. Monopolistic competition under Kimball (1995) demand, in turn, gives rise to endogenously varying markups. Firms are heterogeneous in multiple dimensions, including productivity. Competition in labor and product markets determines the equilibrium dispersion in markdowns and markups and the entry and exit behavior of firms. The model nests numerous heterogeneous firm models with imperfect competition and provides a tractable framework to study equilibrium distortions stemming from monopsony and how they interact with imperfect competition in product markets.

[1] See Manning (2021) and Ashenfelter *et al.* (2021) for recent reviews of the empirical evidence on monopsony.

[2] I show that a homothetic labor supply system with variable elasticities can be generated by aggregating random discrete choices of individual workers. Specifically, I adapt the arguments in Thisse & Ushev (2016) to show that a generalized Kimball (1995) labor supply system, introduced as generalized Kimball demand by Matsuyama & Ushchev (2017), can be microfounded through a multinomial logit choice model that allows for violations of the indpendence of irrelevant alternatives axiom.

In the model, two externalities arise from monopsonistic competition between firms. First, workers earn a surplus from ongoing employment relationships, causing a production externality as firms do not internalize the worker surplus that they create. Firm rents in labor markets, in turn, arise from the profits generated through the suppression of wages, leading to a second production externality as firms do not internalize the effect of their actions on the production incentives of other firms.[3] The magnitude of these externalities varies endogenously across firms and is captured by elasticities related to labor supply. Similarly, markup pricing leads to externalities in product markets that are captured by elasticities related to product demand.

To understand when monopsonistic competition between firms leads to inefficiencies, I compare the market allocation to that chosen by a social planner who is subject to the same technological constraints. I show that homogeneous markdowns are necessary for monopsonistic competition between firms to yield socially optimal market allocations. Further, the decentralized equilibrium is efficient if, and only if, both markdowns and markups are homogeneous across firms. Intuitively, when labor market power is constant across firms, so is the magnitude of the associated externalities. Thus, monopsonistic competition is not a cause of misalignment between social and private production incentives. Homogeneity in market power across firms within all markets uniquely ensures that the externalities from imperfect competition exactly offset each other. In this case, the market induces socially optimal allocations, and laissez-faire policy is optimal.

In contrast, micro-level heterogeneity in markdowns results in inefficient market allocations. I show that the model allows characterizing the ensuing inefficiencies in terms of firm-level elasticities capturing the magnitude of rents arising in labor and product markets. Given an allocation, knowledge of these elasticities allows assessing how monopsonistic competition between firms manifests in aggregate distortions via the allocation of resources across firms, aggregate entry, and selection. For example, pair-wise comparisons of markdowns and markups allow determining whether marginally reallocating labor from one firm to another improves welfare. Similarly, the efficiency of entry is linked to aggregates measuring the respective magnitude of household and firm rents arising in labor and product markets.

An essential practical insight generated by these considerations is that policy prescriptions based on the analysis of markups or markdowns alone may have unintended adverse effects in an economy where firms exercise labor and product market power. To illustrate this point, I use the model to analyze the aggregate impact of two policy interventions frequently emphasized as potential remedies to the adverse effects of imperfect competition. First, I ask when firm-level taxes can successfully restore efficiency. Second, I study the welfare effects of market expansion, i.e., market integration.

[3] This externality is closely related to the "business stealing" externality caused by markups, as described by Mankiw & Whinston (1986).

A robust finding in the literature is that high markup firms should be subsidized. In contrast, I show that in an economy with micro-level heterogeneity in markups and markdowns, markup-based taxation schemes may result in welfare losses. Intuitively, when markups paint an inaccurate picture of the extent to which production decisions are distorted across firms, then policies that reallocate resources towards high markup firms might amplify misallocations stemming from monopsony.

A policy intervention that does not require detailed knowledge of firm-level distortions is market integration, i.e., trade. To establish a benchmark, I first characterize the gains from trade in an economy with homogeneous markdowns and markups, showing that market integration always increases real income and welfare. Quantitative evaluations of the associated gains suggest that real income rises substantially more than in an economy with perfectly competitive labor markets. Intuitively, under monopsony, households earn an employment surplus from each additional entrant. When labor supply is elastic, the increase in real wages following the increase in entry causes aggregate labor supply to rise, which enables further entry.

Meanwhile, in an economy with micro-level heterogeneity in markdowns and markups, gains from market integration are not guaranteed. I use the model to establish sufficient conditions in terms of labor supply and product demand primitives that ensure positive gains from market integration. These conditions generalize the notion of aligned preferences introduced by Dhingra & Morrow (2019). Gains from market integration are guaranteed to be positive when private gains (decreasing in markdowns and increasing in markups) and social gains (household surpluses in product and labor markets) move in the same direction as firm productivity increases.

I then ask when market integration results in the reduction of aggregate distortions. To do so, I follow the conceptual approach in Baqaee & Farhi (2020) and provide a complete decomposition of the welfare changes associated with market expansion in terms of "technical" and "allocative gains." My results highlight that when firms adjust to rising competition by changing prices and wages, responses of aggregate misallocation to market integration might change dramatically, compared to those predicted by benchmark models with competitive labor markets. For example, monopsony might magnify or entirely undo any pro-competitive effects of market integration on markdowns and markups. Similarly, it can impede the beneficial reallocation of resources from low to high markup firms that recent work has stressed as the primary source of aggregate returns to scale. In this context, the model makes two contributions. First, it allows establishing sufficient conditions in terms of model primitives that ensure reallocation gains. When these conditions are met, the gains in an economy with homogeneous markdowns and markups provide a lower bound for the overall gains from market expansion. Second, my results highlight firm-level elasticities that can be used when appropriately aggregated to calculate counterfactual changes in welfare and real income in the model.

The last part of the paper considers extensions of the basic framework. First, I extend the model to account for geographically segmented labor markets with local entry and selection of employers. Second, I introduce heterogeneous worker groups to investigate how heterogeneity in workers' skills or occupations affects potential misallocations from monopsony. In both cases, the efficiency of the decentralized equilibrium and the nature of potential misallocations remain tied to firm-level elasticities of labor supply and product demand. Third, I show that the characterization of efficient market allocations remains unchanged when considering alternative labor supply models with variable elasticities and discuss extending the model to account for multiple industries.

The results in this paper highlight that imperfect competition in labor markets is a potentially important determinant of the welfare implications of firm heterogeneity. How labor market power varies across firms (and labor markets) is key for misallocations, the welfare gains from market expansion, and interacts with imperfect competition in output markets in significant ways. The proposed framework is parsimonious in that it generalizes canonical models of monopolistic competition. Moreover, it is well-suited for future quantitative explorations of the theoretical channels highlighted in this paper.

Related Literature This paper contributes to a large literature investigating the role of firms in shaping aggregate productivity, competition, entry, and welfare. Early analysis of the homogeneous firms case are provided by the seminal work of Spence (1976), Dixit & Stiglitz (1977), Krugman (1979), Venables (1985), and Mankiw & Whinston (1986). Building on this early work, Melitz (2003), Melitz & Ottaviano (2008), Epifani & Gancia (2011), Zhelobodko *et al.* (2012), Melitz & Redding (2015), Mrázová & Neary (2017, 2019), Edmond *et al.* (2018), Arkolakis *et al.* (2019), Bilbiie *et al.* (2019), Dhingra & Morrow (2019), Matsuyama & Ushchev (2020), and Baqaee *et al.* (2021) analyze the implications of firm heterogeneity. The primary focus of this literature is on understanding misallocation driven by markups. In contrast, I focus on misallocations from imperfect competition in labor markets and heterogeneous markdowns. I show that an integrated framework featuring varying markdowns and markups yields new insights into the positive and normative implications of monopolistic competition.

Further, the paper contributes to a large literature exploring the implications of labor market power and monopsony. The model captures key empirical features of firm-level wages documented by the empirical literature estimating firm-level labor supply elasticities and wage pass-throughs (e.g., Staiger *et al.* (2010),Webber (2015), Serrato & Zidar (2016), Garin & Silvero (2018), Dube *et al.* (2020), Bachmann *et al.* (2020)). I provide new insights into how cross-sectional reduced-form estimates can be used to inform aggregate distortions from monopsony. Further, this paper relates to recent work by Brooks *et al.* (2021) and Hershbein *et al.* (2022), who extend the cost based approach to measuring markups (Hall (1988), Loecker & Warzynski (2012)) to measure both markdowns and

markups at the firm-level.

On the theoretical side, my work relates to benchmark models of monopsony described, e.g., in Burdett & Mortensen (1998), Manning (2003), Card *et al.* (2018), Trottner (2020), Haanwinckel (2021), Jha & Rodriguez-Lopez (2021), Lamadon *et al.* (2022), and Kroft *et al.* (2020)). Many of these papers model monopsony as arising from preference heterogeneity over workplace amenities on the worker side. This microfoundation leverages the equivalence of demand systems generated by logit discrete choice and CES utility (Anderson *et al.* (1988)), and provides a tractable CES labor supply system. Building on Thisse & Ushev (2016), I show this approach can microfound richer yet, equally tractable functional forms, such as the Kimball (1995) labor supply system used in the main text. This allows to tractably model endogenously varying markdowns and wage pass-throughs even when firms are atomistic within an economy featuring endogenous entry, exit, and variable markups.

An alternative approach to modeling varying markdowns is to allow for strategic interactions in wage-setting. Recent work in this spirit includes, e.g., MacKenzie (2018), and Berger *et al.* (2022a,b), who extend the quantitative tools introduced by Atkeson & Burstein (2008) to a labor market setting. One key advantage of this approach is that it can credibly account for changes in concentration. A disadvantage is that modeling entry with strategic complementarities is notoriously difficult. Related to these papers focusing on non-atomistic firms, Jarosch *et al.* (2019) study labor market power of granular employers arising in a search setting.

Structure of the paper Section 2 lays out the theoretical framework. Section 3 lays out the social planner's problem and characterizes efficient market equilibria. Section 4 analyzes the policy implications of the model. Section 5 discusses theoretical extensions. Section 6 concludes. All proofs are relegated to Appendix B.

2 Theoretical Framework

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

2.1 Households

Preferences The is populated by a mass L of identical households. Each household derives utility from consumption and labor. Utility is separable in consumption C and labor N :

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

The utility index U is twice continously differentiable and satisfies standard properties.[4] Households consume final goods $\theta \in \Theta$ and supply labor to employers $\omega \in \Omega \equiv \Theta \bigcup \{e, o\}$ producing final, entry, and overhead goods.[5] The consumption index C and the labor supply index N are defined as:

$$1 = \int_{\theta \in \Theta} \Upsilon_\theta(\frac{c_\theta}{C}) dM^C(\theta), \tag{2}$$

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

where c_θ denotes per-capita consumption of good θ, and n_ω denotes per-capita labor supplied to empoyer ω. $dM^C(\theta)$ denotes the mass of a consumption variety θ, while $dM^E(\omega)$ denotes the mass of employers of type ω. Both will be explained further once I introduce the production side of the economy. The consumption utility indices $\Upsilon_\theta(.)$ are strictly increasing and concave and satisfy $\Upsilon_\theta(0) = 0$. The labor disutility indices $\Psi_\omega(.)$ are strictly increasing and convex, and satisfy $\Psi_\omega(0) = 0$.

The functional form of the aggregators in (2) and (3) was introduced by Matsuyama & Ushchev (2017) as a generalization of Kimball (1995) homothetic preferences. CES preferences over consumption of varities and labor supply across employers are a special case of the above aggregators when $\Upsilon_\theta(x) = a_\theta x^{\frac{\sigma-1}{\sigma}}$ and $\Psi_\omega(x) = b_\omega x^{\frac{\beta+1}{\beta}}$ with constants a_θ and b_ω capturing firm-specific taste shifters in consumption and labor supply.

The preferences over labor supply to individual employers imply that jobs are differentiated from the perspective of workers, which will constitute the source of labor market power in the model. A microfoundation is discussed further below.

Utility Maximization Consumers maximize utility $\mathcal{U}$ subject to the following budget constraint,

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

where p_θ is the price of good θ and w_ω is the wage offered by employer ω. The expression for the budget constraint anticipates that there is no aggregate profits redistributed in equilibrium due to free entry, so per-capita wage earnings are equal to per-capita nominal GDP, which is used as the numeraire.

Solving the household's problem,[6] the per-capita inverse demand for product θ is given by,

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

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

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

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

while per-capita labor supply to employer ω equals:

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

$\mathcal{P}$ and $\mathcal{W}$ are price and wage aggregates given by:[7]

$$\mathcal{P} \equiv \frac{\bar{P}}{C}, \qquad \frac{1}{\bar{P}} \equiv \int \Upsilon'_\theta(\frac{c_\theta}{C})\frac{c_\theta}{C} dM^C(\theta), \tag{6}$$

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

It is worth noting that $\mathcal{P}$ and $\mathcal{W}$,in general, do not coincide with the price and wage indices solving the nested expenditure minimization and income maximization problems.[8] Equations (4) and (5) illustrate the appeal of the generalized Kimball aggregators: By choosing Ψ_θ and Υ_ω appropriately, one can generate product demand and labor supply curves to individual firms of any desired shape. Further, firms are allowed to face different residual product demand and labor supply curves, permitting the model to account for, e.g., exogenous differences in non-wage amenities affecting the utility workers obtain from jobs.

$\mathcal{P}$ and $\mathcal{W}$, respectively, mediate monopolistic and monopsonistic competition between firms. The price elasticity of product demand is given by,

$$\sigma_\theta(\frac{c_\theta}{C}) \equiv -\frac{\partial \log c_\theta}{\partial \log p_\theta} = -\frac{\Upsilon'_\theta(\frac{c_\theta}{C})}{\Upsilon''_\theta(\frac{c_\theta}{C})\frac{c_\theta}{C}}, \tag{8}$$

while the wage elasticity of the labor supply to an individual firm is given by

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

The firm-level labor supply elasticity will capture a firm's labor market power. In general, provided that the indices Ψ_ω have varying elasticities, labor market power will endogenously vary if firms offer different wages. In this case, equation (9) shows that exposure to aggregate competition in the labor market also varies across firms. It is worth noting that the model allows labor market power to vary exogenously across firms. Specifically,

[7] These aggregates are, in general, not equal to the ideal price index P and wage index W.

[8] Specifically, let $\mathcal{P}^I$ and $\mathcal{W}^I$ denote the wage indices that satisfy the expenditure and income functions $e(\{p_\omega\}, C) = \mathcal{P}^I C$, and $I(\{w_{\omega'}\}, N) = \mathcal{W}^I N$. Due to the separability of $\mathcal{U}$ and the homotheticity of the aggregators (2) and (3), the household's consumption leisure choice is equivalent to maximizing $\mathcal{U}(C, N)$ subject to $\mathcal{P}^I C = \mathcal{W}^I N$. $\mathcal{P}$ and $\mathcal{W}$ only coincide with $\mathcal{P}^I$ and $\mathcal{W}^I$ when Υ_ω and $\Psi_{\omega'}$ are isoelastic with common elasticity for all ω and ω' respectively. To see this, note that, by definition, $d\log\frac{C}{N} = d\log\mathcal{W}^I - d\log\mathcal{P}^I$, while $d\log\frac{\mathcal{W}}{\mathcal{P}} = d\log\frac{C}{N} + d\log\bar{W} - d\log\bar{P}$. Only when $\frac{\partial\log\Psi_\omega(x)}{\partial\log x} \equiv \frac{\beta+1}{\beta}$ and $\frac{\partial\log\Upsilon_\omega(x)}{\partial\log x} \equiv \frac{\sigma-1}{\sigma}$, these indices coincide.

when $\Psi_\omega(x) = b_\omega x^{\frac{\beta_\omega+1}{\beta_\omega}}$, the elasticity of labor supply differs across firms, but is exogenous to equilibrium outcomes such as wages offered by the firm or the aggregate wage index $\mathcal{W}$. This flexibility allows disentangling potential distortions stemming from exogenously varying labor market power from those posed by endogenously varying labor supply elasticities.

Finally, when aggregate labor supply is endogenous, C and N are chosen by setting $-\frac{U_C}{U_N} = \frac{N}{C}$.

Microfoundation The labor disutility index N has the interpretation that, from the perspective of individual workers, jobs are differentiated in the utility that they provide. This vertical differentiation of jobs is the source of firms' labor market power in the model. To support this interpretation, Appendix A.2 provides an explicit microfoundation rationalizing the labor supply system above. I adapt the arguments in Thisse & Ushev (2016) to show that a Kimball-type labor supply system, which nests CES as a special case,[9] can be microfounded through discrete choices of individuals whose disutility of work associated with a given employer depends on the whole set of employers from which the labor supply decision is made. As shown in the appendix, this approach gives rise to choice probabilities that keep the flexibility of the multinomial logit set-up, but that violate the independence of irrelevant alternatives axiom.

2.2 Firms

Two types of firms populate the economy: Firms producing consumption goods, and firms producing goods required for entry and overhead.[10][11]

2.2.1 Final Good Producers

Final good firms produce differentiated varieties and compete monopolistically in product markets and monopsonistically with all firms in the economy in labor markets.

[9] Anderson et al. (1995) show that the demand system generated by a representative consumer with CES utility can be micro-founded through a multinomial logit model of discrete choice.

[10] This market structure is similar to and Bilbiie *et al.* (2019) and in the spirit of Ghironi & Melitz (2005). One interpretation is, therefore, that there is two production sectors: One producing consumption goods, and one producing new firms.

[11] An alternative formulation of the model is that jobs within firms are differentiated from workers' perspectives. That is, jobs in variable production and jobs that produce overhead are not perfect substitutes for workers. Appendix A.3 develops this formulation of the model in more detail, showing it nests the framework in the main text is a special case. It shows that the model remains equally tractable under this alternative formulation, but implies that overhead costs are heterogeneous across firms. Importantly, prices firms pay for overhead goods do not interact with the wages firms offer to variable production workers.

The model of the product market follows Melitz (2003). To enter the market, final good producers purchase a fixed quantity of entry goods f_e at price p_e. Upon entry, final good producers draw their type θ from a continuous cdf $G(\theta)$ with pdf $g(\theta)$. After receiving their draw, firms purchase a fixed quantity of overhead goods f_o at price p_o to start production. Once production is set up, a firm with draw θ hiring n_θ hours of labor produces output y_θ using a linear production technology given by

$$y_\theta = A_\theta n_\theta.$$

Productivity A_θ is increasing in θ. Firms with draw θ face an inverse per-capita labor supply curve given by (5), and an inverse per-capita product demand curve given by (4).

Profit-maximization implies that a firm's optimal price is a markup over its marginal cost mc:

$$p_\theta = \mu_\theta mc_\theta, \tag{10}$$

where the markup satisfies the usual Lerner formula:

$$\mu_\theta(\frac{c_\theta}{C}) = \frac{\sigma_\theta(\frac{c_\theta}{C})}{\sigma_\theta(\frac{c_\theta}{C}) - 1}, \tag{11}$$

where σ_θ is the price elasticity of demand defined in (8). Profit-maximizing wages equal a markdown $\mathcal{M}$ over a firm's marginal revenue product of labor $mrpl$:

$$w_\theta = \mathcal{M}_\theta mrpl_\theta,$$

where the markdown depends inversely on the labor supply elasticity faced by the firm in (9):

$$\mathcal{M}_\theta(\frac{n_\theta}{N}) = \frac{\beta_\theta(\frac{n_\theta}{N})}{\beta_\theta(\frac{n_\theta}{N}) + 1}. \tag{12}$$

Equations (11) and (12) show that when labor supply and product demand elasticities vary across firms, so will firms' desired markups and markdowns. I impose restrictions on Υ_θ and Ψ_θ sufficient to guarantee that marginal profits are strictly decreasing for all final good producers.

The marginal revenue product of labor equals $mrpl_\theta = \frac{p_\theta}{\mu_\theta} A_\theta$, while marginal costs are equal to $mc_\theta \equiv \frac{w_\theta}{\mathcal{M}_\theta A_\theta}$. Thus, wages and prices are related through following relationship:

$$p_\theta = \frac{\mu_\theta}{\mathcal{M}_\theta} \frac{w_\theta}{A_\theta}. \tag{13}$$

Together, (11), (12), and (13) highlight that prices and wages are jointly determined by a firm's desired level of employment and aggregate conditions in labor and product markets.

Final good producers operate if, and only if, operating profits exceed the costs of overhead:

$$Lp_\theta y_\theta \left(1 - \frac{\mathcal{M}_\theta}{\mu_\theta}\right) \geq p_o f_o. \tag{14}$$

I impose that model primitives are such that variable operating profits are strictly increasing in θ.

Assumption 1 $X_\theta \equiv p_\theta y_\theta \left(1 - \frac{\mathcal{M}_\theta}{\mu_\theta}\right)$ *is continuous and strictly increasing in* θ.

The assumption ensures that while firms are heterogeneous along multiple dimensions, firm profits are strictly increasing in firm productivity. Under this assumption, there exists a unique cutoff θ^* such that firms of type $\theta \geq \theta^*$ produce, while firms with draws $\theta < \theta^*$ exit.

Free entry implies that firms enter until the expected operating profits upon entry equal the entry costs:

$$\int_{\theta^*}^{\infty} \left(L(1 - \frac{\mathcal{M}_\theta}{\mu_\theta}) p_\theta c_\theta - p_o f_o\right) dG(\theta) = p_e f_e. \tag{15}$$

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

2.2.2 Entry and Overhead Good Producers

The output market for overhead and entry goods is perfectly competitive. Goods required for entry and overhead are produced by homogeneous firms using linear production technologies: $y_k = n_k$ for $k \in \{o, e\}$. There is free entry into each output market, and producers compete for workers in the same labor market as final goods producers.

Perfect competititon, free entry, and convex costs together imply that each entry good is provided by a single producer. Each producer charges the same price p_e that equals the average cost of hiring a total f_e hours of labor:

$$p_e = \mathcal{W}\Psi_e'(\frac{f_e}{LN}). \tag{16}$$

The mass of entry good producers, in turn, equals the mass of entrants in the final goods market $dM^E(e) = M$.

Similar arguments imply that there is a mass $dM^E(o) = M[1 - G(\theta^*)]$ of overhead goods producers, each charging price p_o given by,

$$p_o = \mathcal{W}\Psi_o'(\frac{f_o}{LN}). \tag{17}$$

2.3 Equilibrium

Given L, f_e, f_o and $G(.)$, a decentralized equilibrium is a defined by a number of entrants M, an exit cutoff θ^*, allocations $\{n_o, n_e, \{c_\theta, n_\theta\}_\theta\}$, and prices $\left\{p_o, p_e, \{p_\theta, w_\theta\}_\theta\right\}$ such that consumers maximize utility taking prices and wages as given, firms in the final, entry, and overhead goods sector maximize profits, taking the aggregates $C, N, \mathcal{P}$ and $\mathcal{W}$ as given, and markets clear.

2.4 Special Cases & Extensions

The framework nests a number of canonical heterogeneous firm models as special cases, some of which I highlight below.

1. When labor markets are perfectly competitive ($\beta_\omega \to \infty$), N is fixed, $\Upsilon_\theta(x) \equiv x^{(\sigma-1)/\sigma}$, and $A_\theta = \theta$, then the model reduces to the heterogeneous firm model with monopolistic competition introduced by Melitz (2003).

2. When labor markets are perfectly competitive ($\beta_\omega \to \infty$) and N is fixed, the model reduces to (a static version of) benchmark models of monopolistic competition with variable markups analyzed in, e.g., Edmond *et al.* (2018) and Baqaee *et al.* (2021).

3. When N is fixed, and $\Psi_\theta(x) = a_\theta x^{(\beta+1)/\beta}$, then the labor supply system is equivalent to models of static monopsony based on logit discrete-choice described in, e.g. Manning (2011), Card *et al.* (2018), Haanwinckel (2021), Trottner (2020) Lamadon *et al.* (2022), and Kroft *et al.* (2020).

4. Appendix C discusses various extensions of the framework. The model readily extends to incorporate segmented regional labor markets with local entry and exit of firms, as well as heterogeneous worker groups, e.g. skill or occupation.

2.5 Notation

Throughout the rest of this paper, denote the final good sales-weighted average of a variable z by $\mathbb{E}_{pc}[z] \equiv \int_{\theta^*}^{\infty} \omega_{p_\theta c_\theta} z_\theta x_\theta dG(\theta)$ where $\omega_{p_\theta c_\theta} = \frac{p_\theta c_\theta}{\int_{\theta^*}^{\infty} p_{\theta'} c_{\theta'} dG(\theta')}$. Further, denote the wage-bill-weighted average of a variable z by $\mathbb{E}_{wn}[z] = \frac{z_e \times w_e f_e + \int_{\theta^*}^{\infty} \left[z_o w_o f_o + z_\theta w_\theta n_\theta\right] dG(\theta)}{f_e w_e + \int_{\theta^*}^{\infty} \left[w_o f_o + w_\theta n_\theta\right] dG(\theta)}$. Lastly, let $\text{COV}_x[y_\theta, z_\theta] = \mathbb{E}_x[z_\theta y_\theta] - \mathbb{E}_x[z_\theta]\mathbb{E}_x[y_\theta]$.

3 Efficiency

In this section, I highlight firm-level elasticities that are key to understanding aggregate misallocations. I then introduce a benevolent social planner to characterize conditions under which the decentralized equilibrium outlined in Section 2 is constrained-efficient.

3.1 Household rents

Two firm-level parameters are key to understanding the welfare implications of the model. First, the infra-marginal employment surplus of a job ω δ_ω is defined as the labor equivalent disutility from a marginal employer:

$$\delta_\omega \equiv \frac{\Psi(\frac{n_{\theta'}}{N})}{\Psi'(\frac{n_{\theta'}}{N})\frac{n_{\theta'}}{N}} \in (0,1]. \tag{18}$$

The right panel in Figure 3.1 graphically illustrates δ_ω. In a perfectly competitive labor market employers are perfectly substitutable from the perspective of workers, the infra-marginal employment surplus is equal to 1, and earnings exactly compensate workers for the disutility they incur from labor. In an imperfectly competitive labor market, δ_ω is less than 1 and $(1-\delta_\omega)w_\omega n_\omega$ captures total rents workers earn from the ongoing employment relationship with employer ω.[12] The microfoundation of the labor supply system discussed earlier permits the interpretation that worker rents arise from information frictions - firms do not observe workers' idiosyncratic workplace preferences, and, thus, cannot wage discriminate. The size of worker rents relative to total earnings varies endogenously across firms. Further, $1-\delta_\omega$ serves to quantify an approbriability externality arising in labor markets from the fact that firms do not internalize the employment surplus that they generate for their workers.

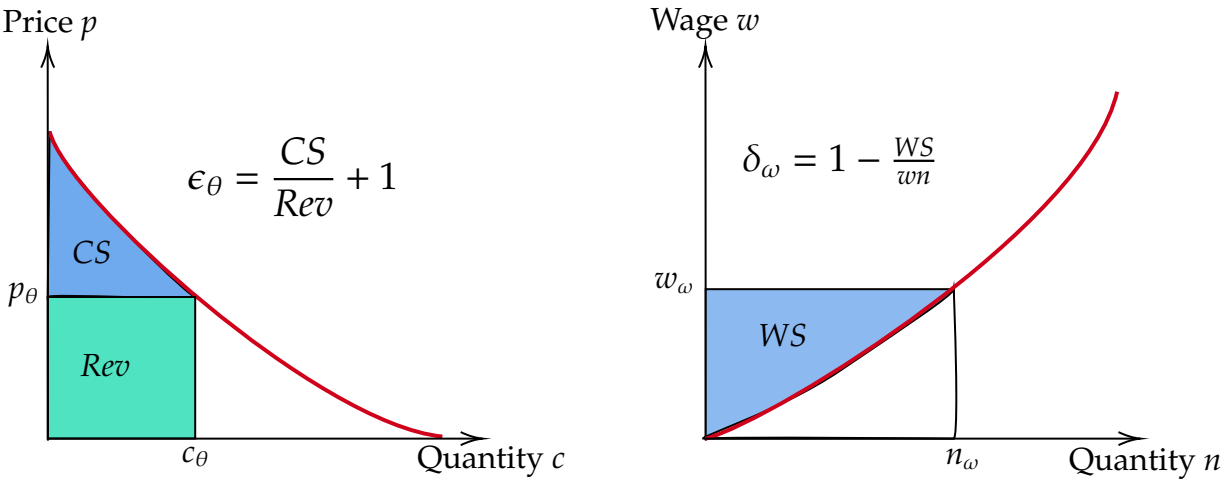
Second, the infra-marginal consumption surplus ϵ_θ associated with consumption variety θ is defined as,

$$\epsilon_\theta = \frac{\Upsilon_\theta(\frac{c_\theta}{C})}{\Upsilon'_\theta(\frac{c_\theta}{C})\frac{c_\theta}{C}} \geq 1. \tag{19}$$

ϵ_θ equals 1 plus the ratio of consumer surplus to revenues, as shown in the left panel of Figure 3.1. Rents earned by households in product markets are, thus, given by $(\epsilon_\theta - 1)p_\theta c_\theta$. Again, the size of rents in product markets relative to consumption spending varies endogenously across varieties. Thereby, ϵ_θ can be thought of to quantify the extent to which firms are unable to capture the entire consumption surplus of their

[12] This definition is conceptually equivalent to the notion of worker rents used in recent work by Lamadon *et al.* (2022) and Kroft *et al.* (2020). In their framework, rents corrspond to a constant fraction of earnings, owing to the fact that labor supply curves are isoelastic, and rents per-employee of a given firm are given by $\frac{1}{\beta_{\omega'}+1}w_{\omega'}$, which is equal to $(1-\delta_{\omega'})w_{\omega'}$ in the special case where $\Psi_{\omega'}(x) = b_{\omega'}x^{\frac{\beta_{\omega'}+1}{\beta_{\omega'}}}$.

Figure 3.1 Consumption and Employment surpluses



production. As in Dhingra & Morrow (2019), rents earned by consumers, thereby, induce an appropriability externality in product markets.

For the remainder of this paper, denote $\bar{\epsilon} = \mathbb{E}_{pc}[\epsilon_\theta]$ and $\bar{\delta}=\mathbb{E}_{wn}[\delta]$. Note that $\bar{\epsilon}$ equals the product demand index $\bar{P}$ defined in (6), while $\bar{\delta}$ equals the labor supply index $\bar{W}$ defined in equation (7).

3.2 The planner's problem

The planner's objective is to maximize household 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 productivity threshold. Therefore, optimal allocation decisions can be summarized by a number of entrants, an exit cutoff, and the labor allocation to and quantities produced by each firm. Formally, the planner solves the following problem:

$$\max_{c_\theta, n_\theta, \theta^*, M_e, n_o, n_e} U(C, N) \tag{20}$$

subject to:

$$1 = M_e \int_{\theta^*}^{\infty} \Upsilon_\theta(\frac{c_\theta}{C}) dG(\theta), \tag{21}$$

$$1 = M_e \left(\Psi_e(n_e/N) + \int_{\theta^*}^{\infty} \left\{ \Psi_o(n_o/N)) + \Psi_\theta(\frac{n_\theta}{N}) \right\} dG(\theta) \right), \tag{22}$$

$$c_\theta \leq n_\theta / A_\theta, Ln_e \geq f_e, Ln_o \geq f_o \tag{23}$$

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

3.3 Inelastic Aggregate Labor Supply

The following theorem provides necessary and sufficient conditions for the equilibrium to be efficient in the case when the aggregate per-capita supply of labor is fixed.[13][14]

Theorem 1. *Suppose that aggregate labor supply is fixed, $N \equiv \bar{N}$. Then, the market allocation 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 $\sigma, \beta > 1$, and $a_\theta, b_\omega \in \mathbb{R}^+$.*

Proof. See Appendix B.1. □

Theorem 1 shows that monopsonistic competition induces socially optimal market allocations if, and only if, markdowns and markups are homogeneous. This highlights that monopsony power in itself is not a distortion, generalizing insights in the monopolistic competition literature (e.g., Bilbiie *et al.* (2019); Dhingra & Morrow (2019)). To build intuition for this result, consider a firm-specific explanation. To this end, the previously defined measures related to household's consumption rents ϵ_θ and worker rents δ_θ can be used to define a firms' "social profit margin" $\epsilon_\theta/\delta_\theta$. The characterization of the planner's problem implies that socially optimal labor allocations of labor across final good firms solve:

$$\Upsilon_\theta\left(\frac{c_\theta^{\text{opt}}}{C^{\text{opt}}}\right) = \frac{\epsilon_\theta}{\delta_\theta}\Psi_\theta\left(\frac{n_\theta^{\text{opt}}}{\bar{N}}\right). \tag{24}$$

Meanwhile, market allocations to a firm θ maximize private profits. The corresponding optimality condition implies that market allocations satisfy the following condition:[15]

$$\Upsilon_\theta\left(\frac{c_\theta^{\text{mkt}}}{C^{\text{mkt}}}\right) = \frac{\mu_\theta}{\mathcal{M}_\theta}\Psi_\theta\left(\frac{n_\theta^{\text{mkt}}}{\bar{N}}\right)\frac{\epsilon_\theta \mathbb{E}_{wn}[\delta]}{\delta_\theta \mathbb{E}_{pc}[\epsilon_\theta]}. \tag{25}$$

When households preferences imply homogeneous markdowns and markups, then the private profit margin $\mu_\theta/\mathcal{M}_\theta$ exactly coincides with the "social profit margin" $\epsilon_\theta/\delta_\theta$, and competition aligns social and private production incentives. When private and social incentives for production coincide, then the market incentivizes exactly the right firms to produce. The proof formalizes this intuition, showing that since private and social profit profits are the aligned, market entry and exit will also be optimal. When entry and exit are optimal, the competition between firms in product and labor markets aligns the price and wage indices to ensure optimal firm-level quantities.

[13] That is household's labor supply decisions to individual firms solve: $\max \int w_{\omega'} n_{\omega'} d\omega$ subject to $1 = \int_{\omega'} \Psi'(\frac{n_{\omega'}}{\bar{N}}) d\omega'$. The equilibrium characterization in Section 2 remains unchanged, except that the condition pinning down C and N, $U_C C = -U_N N$ is no longer needed.

[14] The theorem assumes that an equilibrium exists. Sufficient conditions for equilibrium existence are given in (4).

[15] 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 $C\mathcal{P} = \mathbb{E}_{pc}[\epsilon_\theta]$ and $N\mathcal{W} = \mathbb{E}_{wn}[\delta]$.

Theorem 1 shows that the efficiency of the market equilibrium is tied to isoelastic product demand and labor supply, implying that any micro-level heterogeneity in markdowns (or markups) results in distortions. I turn to developing intuition for the nature of these distortions next.

Characterizing margins of misallocation

When a given allocation is inefficient, welfare can be increased by reallocating labor between entry, overhead, and final good production, while keeping the total amount of labor fixed. In what follows, I define local efficiency for each potential margin of distortion: Relative firm sizes, entry, and exit. These margins are the ones that are relevant for reallocation effects operating in the decentralized equilibrium and informative to a policy-maker seeking to understand the nature of distortions in an observed equilibrium. The following thought experiments consider feasible allocations changes along each of these margins.[16]

To characterize distortions in the relative employment of firms in a given allocation, consider a reallocation of workers from firms in $(\theta', \theta' + d\theta')$ to firms in $(\theta, \theta + d\theta')$ that leaves aggregate labor supply N unchanged. If this reallocation raises welfare, firm θ is too large compared to θ'. The following lemma summarizes when this reallocation is beneficial.

Lemma 1. *In a given allocation, reallocating labor from θ' to θ raises welfare if, and only if, $\frac{\mu_\theta}{\mathcal{M}_\theta} > \frac{\mu_{\theta'}}{\mathcal{M}_{\theta'}}$.*

Proof. See Appendix B.2. □

Lemma 1 provides two important insights. First, pair-wise comparisons of relative firm sizes can be conducted via markdowns and markups alone and do not require knowledge of surpluses earned by workers. Second, distortions in the relative sizes of firms depend on their respective degree of effective market power and not inherently on productivity, the wages that firms pay, the non-wage amenities that they offer to workers, or the prices they charge. As a result, it is not necessarily desirable to reallocate resources to more productive firms or to firms paying the highest wages. This generalizes insights obtained in the context of monopolistic competition described in, e.g., Baqaee *et al.* (2021), cautioning, however, that in an economy with micro-level heterogeneity in markdowns, markup-based comparisons are insufficient to assess cross-sectional misallocations across firms.

To characterize entry efficiency, consider a reallocation that moves workers from variable production to entry and overhead, keeping constant the relative sizes of final good sector

[16] The comparative statics are not equilibrium allocations. The next section will evaluate policies in general equilibrium.

firms, the selection cutoff, and aggregate labor supply. If this reallocation raises welfare, entry is said to be insufficient. Else, it is said to be excessive.

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

$$\frac{\bar{\epsilon}}{\bar{\delta}} > \mathbb{E}_{pc}\left[\frac{\mathcal{M}_\theta}{\mu_\theta}\right]^{-1}. \tag{26}$$

where $\bar{\epsilon} = \mathbb{E}_{pc}[\epsilon_\theta]$ *and* $\bar{\delta}$=$\mathbb{E}_{wn}[\delta]$. *Else, entry is excessive.*

Proof. See Appendix B.3. □

The lemma highlights aggregates of firm-level elasticities that determine whether entry in a given allocation is excessive or insufficient. Recall that $\bar{\epsilon} - 1$ captures the average inframarginal surplus households receive in product markets, while $1-\bar{\delta}$ captures average worker rents arising in labor markets. The left-hand-side of (26) captures the increase in household rents that can be generated by moving labor out of variable production to produce entry and overhead. The lemma shows that higher household rents tend to discourage and, thus, lead to insufficient entry. In constrast, the right-hand-side of (26) describes the cost of generating these rents. It captures firm incentives for entry which are shaped by expected profit margins. Consistent with Theorem 1, when markdowns and markups are constant across firms, the benefits and cost exactly counteract each other, (26) holds with equality, and entry is efficient.[17]

It is worth emphasizing that efficiency requires producers of entry and overhead goods to have labor market power. To see this, consider an economy where markups and markdowns are homogeneous across final good producers and entry and overhead good producers pay competitive wages. Equation (26) implies that entry would be excessive in this economy. Intuitively, wages in this economy 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.

Turning towards selection, consider a marginal increase in the selection cutoff, reallocating the labor freed up by the exiting varieties proportionally to entry, overhead, and variable production. If this reallocation increases welfare, I say that selection is too weak. Else, equilibrium selection is too strong.

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

$$\bar{\epsilon} - \epsilon_{\theta^*} + \frac{\mathcal{M}_{\theta^*}}{\mu_{\theta^*}}(\delta_{\theta^*} - \bar{\delta}) + (1 - \frac{\mathcal{M}_{\theta^*}}{\mu_{\theta^*}})(\delta_o - \bar{\delta}) > 0, \tag{27}$$

[17] Also note that when labor markets are perfectly competititve, (26) simplifies to comparing the inframarginal surplus to a sales-weighted harmonic average of markups, $\bar{\epsilon} > \mathbb{E}_{pc}\left[\frac{1}{\mu_\theta}\right]^{-1}$, as in Baqaee *et al.* (2021).

where $\bar{\epsilon} = \mathbb{E}_{pc}[\epsilon_\theta]$ *and* $\bar{\delta}$=$\mathbb{E}_{wn}[\delta]$. *In the reverse case, selection is too strong.*

Proof. See Appendix B.4. □

Intuitively, whether toughening selection raises welfare, everything else constant, depends on the size of the inframarginal surpluses households obtain via the marginal entrant relative to the rents provided by the average firm in the economy. The first term in (27) captures consumption gains, weighing the surplus generated through the reallocation, $\bar{\epsilon}$, against the surplus provided by the marginal entrant, ϵ_{θ^*}. The second term describes changes in the employment surplus associated with forcing tougher selection. Tougher selection is beneficial to households only when it leads to a positive net change in employment and consumption rents. Again, in the special case of isoelastic labor supply and product demand, (27) holds with equality and selection is efficient.

Together with lemmas 1 and 2, Lemma 3 highlights that the margins of inefficiency resulting from the interplay of monopsonistic and monopolistic competition are locally characterized via firm-level demand and labor supply primitives of the model.[18] These thought experiments cannot inform the aggregate welfare effects associated with different policy interventions by design. However, as will be shown in section 4, the intuitions and insights developed through these thought-experiments are instrumental to understanding how misallcoations adjust in general equilibrium to changes in the economic environment.

I now turn to analyzing the efficiency implications of the model when aggregate labor supply is elastic.

3.4 Elastic Aggregate Labor Supply

Elastic aggregate labor supply introduces an additional source of misallocation by introducing a wedge in the household's labor supply decision. The following result highlights that when markups and markdowns are homogeneous across firms, monopsonistic competition between firms is in itself not a source of inefficiency.

Theorem 2. *Suppose that aggregate labor supply is elastic. If* $\Upsilon_\theta(x) = a_\theta x^{\frac{\sigma-1}{\sigma}}$ *and* $\Psi_\omega(x) = b_\omega x^{\frac{\beta+1}{\beta}}$, *then (a) aggregate labor supply and consumption in the decentralized equilibrium are lower than in the efficient allocation, (b) a leisure tax equal to* $\tau = \frac{\mu}{\mathcal{M}}$ *funded through lump-sum taxes on households induces socially optimal market allocations, and (c) conditional on aggregate labor supply, there is no distortions in entry, exit, and the relative employment allocations across firms.*

Proof. See appendix. □

[18] Appendix C develops a number of extensions of the model, showing that similar characterizations of misallocations apply in economy with geographically segmented labor markets, heterogeneous worker types, or alternative labor supply systems.

The aggregate wedge in a household's aggregate labor supply decisions that appears when labor supply is elastic leads to less than optimal aggregate labor supply and production of consumption goods (part a). When markdowns and markups are homogeneous, one policy instrument is sufficient to correct the resulting distortions (part b). This highlights that firm heterogeneity is not a source of inefficiency, which part (c) further illustrates by showing that marginal changes in the relative firm sizes, entry, or exit that leave aggregate labor supply of a given allocation unchanged cannot improve welfare. Thus, the local characterization of distortions stemming from firm heterogeneity in markups and markdowns remains valid under elastic aggregate labor supply.

For a given market allocation, the efficiency of aggregate labor supply relative to an economy with homogeneous markdowns and markups can be informed through a thought-experiment similar in spirit to the ones considered to characterize the margins of inefficiency in an economy with inelastic labor supply. To this end, consider a reallocation that forces an increase in aggregate labor supply N, allocating the additional hours in a manner that ensures that aggregate profits remain equal to zero (feasibility), the selection cutoff remains unchanged, and relative firm sizes do not change. If this reallocation raises welfare, aggregate labor supply is said to be insufficient, compared to an economy with homogeneous markdowns and markups. Else, it is said to be excessive.

Lemma 4. *Let* $\frac{1}{\mathcal{M}_{o,e}} = \frac{p_e f_e}{p_o f_o + p_e f_e}\frac{1}{\mathcal{M}_e} + \frac{p_o f_o}{p_o f_o + p_e f_e}\frac{1}{\mathcal{M}_o}$. *In a given market allocation, micro-level heterogeneity in markups and markdowns leads to insufficient aggregate labor supply if, and only if,*

$$\frac{\bar{\epsilon} - \bar{\delta}}{\bar{\delta}} > \mathbb{E}_{pc}[1 - \frac{\mathcal{M}_\theta}{\mu_\theta}]\frac{1}{\mathcal{M}_{o,e}}\mathbb{E}_{pc}[\frac{1}{\mu_\theta}]^{-1} \tag{28}$$

In the reverse case, aggregate labor supply is excessive.

Proof. See Appendix B.6. □

It is worth noting that this result does not indicate that labor supply is insufficiently low compared to the first-best allocation. Instead, it indicates whether heterogeneity in markdowns and markups leads to levels of aggregate labor supply that are higher or lower than in an economy with homogeneous markdowns and markups. While the condition stated in the lemma is related to the condition pinning down the efficiency of entry in an economy with inelastic labor supply, one does not necessarily imply the other.

4 Policy implications

When market allocations reflect distortions caused by imperfect competition in labor markets, a natural question is how policy might be used to reduce misallocations. In this

section, I first analyze when firm-level taxation may restore efficiency. Then, I consider the model's implications for the welfare effects of market expansion as an example of a policy that does not require detailed information on firm-level distortions.

4.1 Welfare determinants

The determinants of how per-capita welfare responds to counterfactual changes in the economic environment can be compactly summarized using the equivalent variation. Changes in welfare relative to a given initial allocation are proportional to changes in real income, $\frac{C}{N}$,[19] and given by:

$$
\begin{aligned}
d\log \mathcal{U} \propto d\log \tfrac{C}{N} = \; & \underbrace{\left(\mathbb{E}_{pc}\left[\epsilon_\theta\right] - \mathbb{E}_{wn}\left[\delta\right]\right) d\log M}_{\Delta \text{ Entry}} \\
& + \underbrace{\omega^{pc}_{\theta^*}\left(\mathbb{E}_{pc}\left[\bar{\epsilon} - \epsilon_{\theta^*}\right] - \frac{\mathcal{M}_{\theta^*}}{\mu_{\theta^*}}(\bar{\delta} - \delta_{\theta^*}) - (1 - \frac{\mathcal{M}_{\theta^*}}{\mu_{\theta^*}})(\bar{\delta} - \delta_o)\right)\frac{g(\theta^*)d\theta^*}{1 - G(\theta^*)}}_{\Delta \text{ Exit}}, \\
& + \underbrace{\mathbb{E}_{pc}\left[\left(1 - \frac{\mathcal{M}_\theta}{\mu_\theta}\right) d\log\left(\frac{\mathcal{M}_\theta}{\mu_\theta}\right)\right] + \mathbb{E}_{pc}[1 - \frac{\mathcal{M}_\theta}{\mu_\theta}] d\log w_{o,e}}_{\Delta \text{ in prices and wages}}
\end{aligned}
\tag{29}
$$

where $w_{o,e}$ denotes the average wage paid to workers in the entry and overhead good sectors, and $\mathbb{E}_{pc}[1 - \frac{\mathcal{M}_\theta}{\mu_\theta}]$ is the share of national income allocated to the entry and overhead good sectors in the initial allocation.

As standard in monopolistic competition models, real income changes depend on changes in the equilibrium number of entrants, selection, and the distribution of markups. However, in contrast to models assuming perfectly competitive labor markets, when firms compete monopsonistically in labor markets, changes in welfare, in addition, stem from changes in the distribution of wages, as well as changes in inframarginal employment surpluses. The last term captures welfare effects arising from changes in wages earned in the sector producing entry and overhead goods.

Equation 29 highlights that in the absence of (endogenous) variation in markdowns and markups, all changes in real income and welfare operate through the entry behavior of firms.

[19] Welfare correspond to changes in $\frac{C}{N}$, scaled by the elasticity of utility with respect to consumption: $d\log \mathcal{U} = \frac{\partial \log \mathcal{U}}{\partial \log C} d\log C + \frac{\partial \log \mathcal{U}}{\partial \log N} d\log N = \frac{\partial \log \mathcal{U}}{\partial \log C} d\log \frac{C}{N}$, where the last equality follows from the household's optimality condition for trading off consumption against labor.

4.2 Firm-level Subsidies and Taxes

A robust finding in monopolistic competition models with variable markups is that efficiency can be restored through taxes and subsidies that incentivize firms with high degrees of market power to increase production. I illustrate that such policies can be detrimental to welfare in the presence of monopsony.

To simplify the analysis, I consider a version of the economy described in Section 2 with exogenous, but heterogeneous markdowns and markups.[20] I first derive the set of welfare-restoring taxes and subsidies in this environment.

Proposition 1. *Suppose that* $\Psi_{\theta'}(\frac{n_{\omega'}}{N}) = b_{\theta'}\left(\frac{n_{\omega'}}{N}\right)^{\frac{\beta_{\omega'}+1}{\beta_{\omega'}}}$ *and* $\Upsilon_{\omega}(\frac{c_{\omega}}{C}) = a_{\omega}\left(\frac{c_{\theta}}{C}\right)^{\frac{\sigma_{\omega}-1}{\sigma_{\omega}}}$. *In addition to a tax on leisure, the following schedule of (implicit) taxes on sales of final, entry, and overhead good producers, financed through lump-sum taxes levied on households, restores the constrained-efficient allocation:*

$$\tau_{\theta} = \left(\frac{\mu_{\theta}\mathbb{E}_{wn}\left[\mathcal{M}\right]}{\mathcal{M}_{\theta}\mathbb{E}_{pc}\left[\mu_{\theta}\right]}\right)^{-1}, \quad \tau_{e} = \mathcal{M}_{e}/\mathbb{E}_{wn}\left[\mathcal{M}\right], \quad \tau_{o} = \mathcal{M}_{o}/\mathbb{E}_{wn}[\mathcal{M}]. \tag{30}$$

Proof. See appendix. □

The efficiency-restoring tax schedule subsidizes firms with higher than average overall market power ($\frac{\mu_{\theta}}{\mathcal{M}_{\theta}}$) in the final goods sector, while it taxes firms with lower than average market power. Producers in the entry and overhead goods sector receive subsidies if they have higher than average labor market power and are taxed if they have less than average labor market power. The tax scheme is implicit in that the averages of markdowns and markups required to set the economy-wide markups and markdowns are endogenous to the equilibrium exit behavior of firms.

Proposition 1 highlights that the prevalent policy prescription obtained in economies with variable markups no longer necessarily holds in economies where firms exercise labor market power. In this case, it is natural to ask how well-intended policies based on the analysis of markups affect distortions and welfare. The following result analyzes when markup-based interventions are welfare enhancing.

Proposition 2. *Suppose that* $\Psi_{\omega}(\frac{n_{\omega}}{N}) = b_{\omega}\left(\frac{n_{\omega}}{N}\right)^{\frac{\beta_{\omega}+1}{\beta_{\omega}}}$ *and* $\Upsilon_{\theta}(\frac{c_{\theta}}{C}) = a_{\theta}\left(\frac{c_{\theta}}{C}\right)^{\frac{\sigma_{\theta}-1}{\sigma_{\theta}}}$. *Consider a policy-maker who implements a set of sales taxes and subsidies funded by lump-sum taxes on the household that induces firms to price at a constant markup over marginal cost:* $\tau_{\theta} = \left(\frac{\mu_{\theta}}{\mathbb{E}_{pc}\left[\mu_{\theta}\right]}\right)^{-1}$. *This policy raises welfare if* $\mathcal{M}_{\theta} < \mathcal{M}_{\theta'}$ *whenever* $\mu_{\theta} > \mu_{\theta'}$. *It lowers welfare if* $\frac{\mu_{\theta'}}{\mathcal{M}_{\theta'}} > \frac{\mu_{\theta}}{\mathcal{M}_{\theta}}$ *whenever* $\mu_{\theta} > \mu_{\theta'}$.

Proposition 2 shows that firm-level interventions that ignore monopsony when correcting inefficiencies from imperfect competition may decrease welfare. This reflects the intuition

[20] This allows to ignore issues related to firms internalizing the effect of taxation when choosing optimal markups and markdowns.

that correcting one source of inefficiency may amplify the magnitude of distortions caused by another to the point that the policy causes a loss in welfare larger than the one the policy sought to correct.

The economic reasoning underlying this result is simple: When markups are not sufficient to pairwise rank firms in terms of their markup to markdown ratios, policies that reallocate resources toward high markup firms may have unintended adverse effects by causing further misallocations. Consequently, when markup comparisons are sufficient to rank any pair of firms in terms of their exercised market power in output and input markets, a policy that reallocates resources to high markup firms raises welfare (first part). Conversely, in an economy where markup-based comparisons always lead to incorrect pairwise rankings of firms in terms of their markdown to markup ratios, a policy that redistributes resources to high markdown firms may amplify distortions from monopsony to cause additional welfare losses (second part).

While both cases constitute extreme examples, they serve to illustrate the idea that all sources of inefficiency have to be jointly assessed in order to design welfare-enhancing policy interventions at the firm level. With this in mind, I now turn to analyze how policies that increase aggregate competition affect welfare in the economy.

4.3 Market Expansion

While firm-level policy intervention requires detailed knowledge of how distortions vary across firms, an alternative that has received much attention from policy-makers and academics is integration with world markets. Market integration, loosely speaking, promotes aggregate competition and, thereby, might help alleviate distortions from market power. In the model, market integration can be captured in a stylized way through an increase in the size of the population L.[21] I begin by investigating the gains from market expansion in an economy with homogeneous markdowns and markups. I then provide sufficient conditions in terms of labor supply and product demand primitives that ensure overall positive gains from market expansion. Finally, I provide conditions under which the gains from market expansion are larger than in an economy with homogeneous markdowns and markups.

[21] This goes back to Krugman (1979), who establishes that free trade between countries with the same tastes and technologies yields allocations that mimique those prevailing in an economy with the same combined population.

4.3.1 Gains under homogeneous markdowns and markups

Proposition 3 establishes that under regularity conditions,[22] an increase in market size always raises welfare in an economy with homogeneous markdowns and markups.

Proposition 3. *Consider an economy with constant markdowns* $\mathcal{M} = \bar{\delta}$ *and markups* $\mu = \bar{\epsilon}$ *. Denote* $\phi \equiv \frac{1+\varepsilon_C^{U_C}+\varepsilon_N^{U_C}}{1+\varepsilon_C^{U_N}+\varepsilon_N^{U_N}}$*. If* $\phi < \frac{\mathcal{M}}{\mu}$*, then the equilibrium exists, is unique, and an increase in market size* $d\log L > 0$ *always raises real welfare:*

$$d\log \mathcal{U} \propto d\log \frac{C}{N} = \left(\frac{\mu}{\mathcal{M}} - 1\right) d\log L + \frac{\phi\left(\frac{\mu}{\mathcal{M}} - 1\right)}{1 - \phi - \phi\left(\frac{\mu}{\mathcal{M}} - 1\right)} d\log L. \tag{31}$$

Proof. See Appendix B.7.3. □

Gains can be decomposed into the gains that would occur in an economy with a fixed labor supply and the additional gains generated by endogenous responses in labor supply. As the Frisch elasticity of labor supply goes to zero, $\phi \to 0$ and welfare changes are fully captured by the first term. Further, the gains from market expansion can be calculated without knowledge of the underlying firm type distribution $G(.)$, and are characterized by the economy-wide markdown, markup, and the total elasticity of aggregate labor supply to changes in real wages.

Proposition 3 suggests that the gains from market expansion are likely to be substantially higher in an economy with homogeneous markdowns and endogenous labor supply, compared to the canonical heterogeneous firm model with monopolistic competition and homogeneous markups described in Melitz (2003). To illustrate this, Table 4.1 displays the gains of market expansion in terms of real income under various parametrizations of the economy, assuming preferences over leisure and consumption take the Greenwood *et al.* (1988) form, $\mathcal{U}(C, N) = C + \frac{N^{1+\frac{1}{\varphi}}}{1+\frac{1}{\varphi}}$, so that $\phi = \frac{1}{1+\frac{1}{\varphi}}$. Common values for φ considered in the literature range from 0.2 to 0.8, which translates into values of ϕ ranging from 0.15 to 0.5

Table 4.1 shows that compared to an economy with competitive labor markets and fixed labor supply, the real income gains from market expansion are higher in an economy with monopsonistic labor markets. Real income gains from market expansion more than double in an economy with constant markdowns and endogenous labor supply even under modest calibrations for the size of markdowns and the Frisch elasticity. The combination of two effects explains this result. First, households earn an additional inframarginal employment surplus from each entrant under monopsony. Second, under

[22] The regularity condition ensures equilibrium stability by imposing that changes in income from entry cannot lead to more than proportional increases or decreases in labor supply. Given benchmark empirical estimates, of the Frisch elasticity and markdowns and markups, this condition is likely to be satisfied.

Table 4.1 Gains from Market Expansion under Constant Markdowns and Markups

	$\frac{d \log C/N}{d \log L}$			
	Competitive Labor Market		Monopsony	
Frisch Elasticity	$\mu = 1.1$	$\mu = 1.2$	$\mu = 1.1, \mathcal{M} = 0.9$	$\mu = 1.2, \mathcal{M} = 0.7$
$\phi = 0$	0.1	0.2	0.22	0.71
$\phi = 0.15$	0.13	0.37	0.26	0.85
$\phi = 0.5$	0.17	0.56	0.48	1.62

endogenous labor supply, additional entry raises labor market competition and results in higher wages, which increases labor supply, resulting in more entry. Thus, endogenous labor supply leads to a "multiplier" that magnifies the benefits of market expansion, and the size of the multiplier is increasing in the rents that households earn in labor and product markets.

These results suggest that market expansion is a potentially powerful policy tool to raise welfare. I now describe how these conclusions may change in an economy with micro-level heterogeneity in markdowns and markups.

4.3.2 Gains from Market Expansion under Heterogeneous Markdowns and Markups

Under micro-level heterogeneity in markdowns and markups, gains from market expansion are not guaranteed. This can be illustrated by the model-implied equilibrium comovement of aggregate competition and real income in response to an increase in market size,

$$\underbrace{d \log \frac{\mathcal{W}}{\mathcal{P}}}_{\text{Competition}} = \underbrace{\mathbb{E}_{pc}\left[1 - \frac{\mathcal{M}_\theta}{\mu_\theta}\right] \frac{1}{\mathcal{M}_{o,e}} d \log (NL)}_{\text{Entry}} + \underbrace{\mathbb{E}_{pc}\left[\frac{1}{\sigma_\theta}\right] d \log \left(\frac{C}{N}\right)}_{\text{Reallocation}}. \tag{32}$$

Changes in competition are jointly determined with changes in effective labor supply and real income. First, an increase in market size induces additional entry, which causes changes in aggregate competition. Second, changes in aggregate competition, in turn, cause reallocations between firms, which, in turn, lead to changes in real income. Changes in real income change entry and, thus, competition. In a model with homogeneous markdowns and markups, changes in competition are precisely equivalent to changes in real income, implying that the second effect is absent and that market size always raises real income, as described in Proposition 3. Conversely, in an economy with micro-level heterogeneity in markdowns and markups, changes in real income from market size may

be negative if increased competition sufficiently amplifies misallocations.

Sufficient Conditions for Positive Gains

I begin by generalizing the concept of aligned preferences introduced by Dhingra & Morrow (2019) to an economy with monopsonistic labor markets. Preferences are aligned if private gains (decreasing in markdowns and increasing in markups) and social gains (household surpluses in product and labor markets) move in the same direction as we move along the firm productivity distribution.

Definition 1 *Assume that Assumption 1 holds. Social and private preferences are said to be aligned if* $\frac{\partial(1-\frac{\mathcal{M}_\theta}{\mu_\theta})}{\partial\theta}$ *and* $\frac{\partial(\frac{\epsilon_\theta}{\delta_\theta}-1)}{\partial\theta}$ *are monotonous with the same sign.*

The following proposition shows that positive gains from market expansion can be ensured under aligned preferences. This is true for any underlying distribution of firm types that satisfies Assumption 1, but requires additional regularity conditions to ensure that that marginal profits are decreasing.

Proposition 4. *Market expansion increases welfare when social and private preferences are aligned (provided that (i) for all* $\theta, \sigma_\theta(.) \geq 2$ *and* $\delta_\theta(.) \leq \min\{\delta_o, \delta_e\}$ *when* $\frac{\mathcal{M}_\theta}{\mu_\theta}$ *is increasing in productivity, and (ii)* $\delta_\theta(.) > \min\{\delta_o, \delta_e\}$ *when* $\frac{\mathcal{M}_\theta}{\mu_\theta}$ *is decreasing in productivity).*

Proof. See Appendix B.7.4. □

The economic reasoning underlying this result is that as market size expands, increased entry and competition reduce quantities and employment across all final good producers. Regarding labor allocations between firms, aligned preferences ensure that the associated changes in profit margins across firms correspond to changes in social profit margins. As a result, the increase in competition incentivizes the right set of firms to expand relative to other firms.

The additional conditions imposed on employment surpluses in the overhead and entry good sector help ensure that any change in selection from market expansion is beneficial under aligned preferences. When profit margins increase in productivity A_θ, then an increase in competition toughens selection. In this case, the stated conditions ensure that selection was initially too weak (as defined in Lemma 3), ensuring that tougher selection raises welfare. If, in contrast, profit margins decrease firm productivity, increased competition weakens selection. Aligned preferences ensure that weaker selection is beneficial when profit margins decrease in productivity as the marginal entrant provided greater consumption and employment surpluses to households than the average active final good producer.

It is worth discussing how Proposition 4 relates to existing results in the literature on monopolistic competition. In an economy with perfectly competitive labor markets, Definition 1 reduces to requiring that markups and inframarginal consumption surpluses move in the same direction as firm productivity varies. Helpman & Krugman (1987) show that aligned preferences are sufficient to guarantee welfare gains from market expansion when firms are homogeneous, while Dhingra & Morrow (2019) show that the same is true when firms are heterogeneous. Proposition 4 extends these insights in two dimensions. First, it shows that under perfect competition in the labor market, the same set of conditions ensure gains from market expansion when aggregate labor supply is elastic. Second, Proposition 4 shows that the notion of aligned preferences that ensures gains from market expansion in an economy with endogenously variable markdowns differs quite substantially from an economy with perfectly competitive labor markets. When labor markets are monopsonistic, an analysis of markdowns - or markups - alone is insufficient to ensure positive gains from market expansion.

Decomposition of the aggregate gains from market expansion

Having established sufficient conditions for market expansion to raise real income, I now discuss conditions under which market expansion leads to beneficial reallocations. To this end, I follow the conceptual approach in Baqaee & Farhi (2020) and Baqaee *et al.* (2021) and decompose welfare changes in market expansion into (i) gains that would materialize absent any reallocations across firms, and (ii) gains that materialize through reallocations across firms via various margins. This decomposition serves two purposes. First, it allows establishing sufficient conditions for the gains from market expansion to be larger in an economy with variable markdowns and markups compared to an economy with homogeneous markdowns and markups. When these conditions are met, the real income gains described in Proposition 3 provide a lower bound for the overall gains from market integration. Second, the decomposition highlights firm-level elasticities that are sufficient to calculate counterfactual changes in real income in the model.

To develop intuition for the decomposition, it is helpful to consider the model-implied partial equilibrium response in key firm outcomes to changes in aggregate competition and real income. To this end, additional concepts related to how firms optimally adjust markdowns and markups are helpful. First, the pass-through of shocks to the marginal revenue productivity of labor into wages can be defined in terms of the elasticity of the markdown:

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

Under isoelastic labor supply curves, the pass-through of labor revenue productivity

shocks into wages is complete as firms have constant markdowns, $\gamma_\theta = 1$. In general, desired markdowns vary with firm size, and pass-through is incomplete.

Second, the pass-through of firm-level changes in marginal cost into prices is defined in terms of the elasticity of the markup:

$$\rho_\theta \equiv \frac{\partial \log p_\theta}{\partial \log mc_\theta} = \frac{1}{1 - \frac{\partial \log \mu_\theta(\frac{p_\theta}{\mathcal{P}})}{\partial \log p_\theta}}. \tag{33}$$

Jointly with the labor supply elasticity β_θ and demand elasticity σ_θ, the elasticities ρ_θ and γ_θ fully describe the partial equilibrium response in relative firm sizes to changes in aggregate market conditions. In the model, absent any firm-level productivity shocks, the relative size of firm θ comoves with aggregate competition, $d \log \frac{\mathcal{W}}{\mathcal{P}}$, and real income, $d \log \frac{C}{N}$, as follows:

$$d \log \frac{c_\theta}{C} = -\frac{\sigma_\theta \rho_\theta \beta_\theta \gamma_\theta}{\sigma_\theta \rho_\theta + \beta_\theta \gamma_\theta} d \log \frac{\mathcal{W}}{\mathcal{P}} - \frac{\sigma_\theta \rho_\theta}{\sigma_\theta \rho_\theta + \beta_\theta \gamma_\theta} d \log \frac{C}{N}. \tag{34}$$

Focusing on the role of the labor supply elasticity β_θ the first term shows that firms with more labor market power, ceteris paribus, contract less as aggregate competition increases. Conversely, the second term shows that firms with lower markdowns tend contract relatively more in response to an increase in real income, reflecting that firms facing less elastic labor supply find it harder to expand production and, thus, serve an increase in aggregate demand. This effect is absent in perfectly competitive labor markets models, highlighting that monopsony introduces upward-sloping marginal cost, with elasticities that vary across firms when markdowns are heterogeneous.

Further, equation (34) shows that firms with lower wage pass-through γ_θ contract relatively less in response to rising competition: Firms with lower pass-through have more elastic markdowns, allowing them to respond to increased competition by lowering profit margins and prices, allowing them to retain employment. The model-implied partial equilibrium response of firms' profit margins to changes in aggregate competition and real income illuminates this:

$$d \log \frac{\mu_\theta}{\mathcal{M}_\theta} = -\left(1 - \rho_\theta \gamma_\theta \frac{\beta_\theta + \sigma_\theta}{\sigma_\theta \rho_\theta + \gamma_\theta \beta_\theta}\right) d \log \frac{\mathcal{W}}{\mathcal{P}} - \frac{1}{\beta_\theta}\left(1 - \rho_\theta \frac{\beta_\theta + \sigma_\theta}{\sigma_\theta \rho_\theta + \gamma_\theta \beta_\theta}\right) d \log \frac{C}{N}. \tag{35}$$

In contrast, the second term in (35) shows that firms with lower wage pass-through γ_θ respond to rising aggregate demand by raising profit-margins and increasing wage markdowns. As a result, these firms contract relatively more as aggregate demand rises, which is also implied by equation (34).

Together, these considerations highlight that heterogeneity in markdowns introduces additional forces that may fundamentally change the reallocation effects of market ex-

pansion underscored in the literature on monopolistic competition. Absent monopsony, a sufficient condition for increased competition, and thus market expansion, to have "pro-competitive" effects on aggregate markups is that firms with higher markups μ_θ have lower price pass-through ρ_θ.[23] In contrast, monopsonistic competition between firms may drastically changes these conclusions, both qualitatively and quantitatively.

Going forward, I denote that partial elasticity of relative firm sizes with respect to competition by χ_θ. χ_θ^{FPT} denotes the same elasticity in an economy with complete pass-through,

$$\chi_\theta \equiv -\frac{\partial \log\left(\frac{c_\theta}{C}\right)}{\partial \log \frac{w}{P}} = \frac{\sigma_\theta \rho_\theta \beta_\theta \gamma_\theta}{\sigma_\theta \rho_\theta + \beta_\theta \gamma_\theta}, \; \chi_\theta^{FPT} \equiv -\frac{\partial \log\left(\frac{c_\theta}{C}\right)}{\partial \log \frac{w}{P}} \frac{1}{\chi_\theta} \frac{\sigma_\theta \beta_\theta}{\sigma_\theta + \beta_\theta} = \frac{\sigma_\theta \beta_\theta}{\sigma_\theta + \beta_\theta} \tag{36}$$

Using these definitions, I characterize the gains from market expansion in general equilibrium, explicitly highlighting reallocative effects stemming from changes in entry, selection, and markdowns/markups.

Proposition 5. *Following a change in market size $d \log L$ changes in real income can be computed from $d \log C = \frac{1}{\phi} d \log N$ and:*

$$d \log \frac{C}{N} = \frac{\bar{\epsilon} - \bar{\delta}}{\bar{\delta}} d \log NL + \frac{\zeta_{entry} + \zeta_{\mathcal{M}/\mu} + \zeta_{\theta^*}}{\chi} d \log NL \tag{37}$$

where

$$\zeta_{entry} = COV_{pc}\left[\left(\frac{\bar{\epsilon}}{\bar{\delta}} \frac{\mathcal{M}_\theta}{\mu_\theta} - 1\right) \chi_\theta^{FPT}, \frac{1}{\mu_\theta} - \frac{\bar{\delta}}{\bar{\epsilon}} \mathcal{M}_\theta\right] + \frac{\bar{\delta}}{\bar{\epsilon}} \mathbb{E}_{pc}\left[\left(\frac{\bar{\epsilon}}{\bar{\delta}} - \frac{\mu_\theta}{\mathcal{M}_\theta}\right) \chi_\theta^{FPT}\right] \left[\frac{\mathbb{E}_{pc}[\mathcal{M}_\theta] - \mathcal{M}_{o,e}}{\mathbb{E}_{pc}\left[\frac{1}{\mu_\theta}\right]^{-1}}\right] \tag{38}$$

$$\zeta_{\mathcal{M}/\mu} \equiv \frac{\bar{\delta}}{\bar{\epsilon}} \mathbb{E}_{pc}\left[\left(1 - \frac{\bar{\epsilon}}{\bar{\delta}} \frac{\mathcal{M}_\theta}{\mu_\theta}\right)\left(\chi_\theta^{FPT} - \chi_\theta\right)\right]\left(\mathbb{E}_{pc}\left[\frac{\mu_\theta - \mathcal{M}_\theta}{\mu_\theta \mathcal{M}_{o,e}}\right] - \mathbb{E}_{pc}\left[\frac{1}{\mu_\theta}\right] \frac{\bar{\epsilon} - \bar{\delta}}{\bar{\delta}}\right) + \mathbb{E}_{pc}\left[\left(1 - \frac{\bar{\epsilon}}{\bar{\delta}} \frac{\mathcal{M}_\theta}{\mu_\theta}\right)\left(\chi_\theta^{FPT} - \chi_\theta \frac{1}{\gamma_\theta}\right) \frac{\bar{\epsilon} - \bar{\delta}}{\beta_\theta}\right] \tag{39}$$

$$\zeta_{\theta^*} \equiv \frac{1}{\mathcal{M}_{o,e}} \iota_\theta b_{\theta^*} \left(1 - \mathbb{E}_{pc}\left[\frac{\mu_\theta - \mathcal{M}_\theta}{\mu_\theta}\right] \frac{\mu_{\theta^*}}{\mu_{\theta^*} - \mathcal{M}_{\theta^*}}\right) \frac{\bar{\epsilon} - \bar{\delta}}{\bar{\delta}} \mathbb{E}_{pc}\left[\frac{1}{\mu_\theta}\right] \tag{40}$$

and $\iota_{\theta^*} \equiv \omega_{\theta^*}^{pc}\left(\epsilon_{\theta^*} - \bar{\epsilon} - (1 - \frac{\mathcal{M}_{\theta^*}}{\mu_{\theta^*}})(\delta_{\theta^*} - \bar{\delta}) - \frac{\mathcal{M}_{\theta^*}}{\mu_{\theta^*}}(\delta_o - \bar{\delta})\right)$, χ_θ is defined in (36), and b_{θ^*} is a constant provided in the appendix. Further,

$$\chi \equiv \mathbb{E}\left[\frac{\mathcal{M}_\theta}{\mu_\theta} - \left(\frac{\bar{\delta}}{\bar{\epsilon}} - \frac{\mathcal{M}_\theta}{\mu_\theta}\right) \chi_\theta \left(\frac{1}{\beta_\theta \gamma_\theta} + \mathbb{E}_{pc}\left[\frac{1}{\sigma_\theta}\right]\right)\right] + \iota_{\theta^*} \zeta_{\theta^*} \left(\mathbb{E}_{pc}\left[\frac{1 - \mathcal{M}_\theta}{\mu_\theta}\right]^{-1} - \frac{\mu_{\theta^*}}{\mu_{\theta^*} - \mathcal{M}_{\theta^*}}\right) \mathbb{E}_{pc}\left[\frac{1}{\mu_\theta}\right] \tag{41}$$

Proof. See Appendix B.7.5. □

[23] This is true whenever preferences satisfy Marshall's second law of demand. See, for example, the discussion in Mrázová & Neary (2017) and Arkolakis *et al.* (2019).

In an economy with a common markup μ and markdown $\mathcal{M}$, $\bar{\epsilon} = \mu$ and $\bar{\delta} = \bar{\mathcal{M}}$. Thus, the first term in equation (37) describes gains from market expansion that would occur if resource allocations were held constant across firms. The second term in equation (37), in turn, highlights that reallocation gains materialize via three margins. First, ζ_{entry} describes how heterogeneity in labor demand and product demand elasticities affects the reallocation effects of market expansion operating through entry. $\zeta_{\mathcal{M}/\mu}$ captures reallocative gains driven by changes in markdowns and markups, and ζ_{θ^*} captures gains from market expansion realized through changes in selection. I now discuss each of these terms in greater detail.

Reallocations via heterogeneity in labor supply and product demand elasticities ζ_{entry} ζ_{entry} consists of two components. The first component captures what Baqaee *et al.* (2021) describe as "Darwinian" gains from market expansion in the context of monopolistic competition: As discussed before, an increase in aggregate competition reallocates resources, ceteris paribus, towards firms facing less elastic product demand and labor supply. The covariance formalizes the intuition that if markdown and markups are sufficiently correlated in a given allocation, competition reallocates resources towards the right set of firms. Note that in an economy with competitive labor markets, this component is always positive, independent of the shape of demand or the nature of social preferences. In contrast, this term is not necessarily positive in an economy with imperfectly competitive labor markets.

The second component only appears in an economy with imperfectly competitive labor markets. It captures the allocative benefits stemming from differences in labor market power across sectors. When labor market power is lower among entry-good producers, then Lemma 2 implies that entry is initially insufficient. In this case, market expansion tends to cause a larger decline in prices in the entry goods sector, which, in turn, reallocates resources towards it. This in turn is beneficial, given that entry was initially insufficient. A similar argument can be made to show that when labor market power is higher among entry-good producers, entry was initially excessive, and market expansion tends to reallocate resources away from the entry sector. Thus, the second term always captures beneficial reallocations.

Given these considerations, the following conditions ensure that $\zeta_{\text{entry}} > 0$.

Corollary 1. *Consider the economy described in section 2 with $f_o = 0$, $\Upsilon_\theta(x) = A_\theta x^{\frac{\sigma_\theta - 1}{\sigma_\theta}}$ and $\Psi_{\omega'}(x) = B_{\omega'} x^{\frac{\beta_{\omega'}+1}{\beta_{\omega'}}}$. In this economy, the gains from market expansion are larger than in an economy with constant markups and markdowns if $\frac{\mathcal{M}_\theta}{\mu_\theta}$ is strictly decreasing in productivity.*

Changes in markdowns and markups: $\zeta_{\mathcal{M}/\mu}$ When markdowns and markups vary endogenously across firms, then changes in competition affect the distribution of mark-

downs and markups in the economy. The term $\zeta_{\mathcal{M}/\mu}$ only appears in an economy where wage or price pass-throughs are incomplete. It captures two forces: One familiar from models focusing exclusively on monopolistic competition and one specific to imperfect competition in the labor market. I label the former $\zeta^{comp}_{\mathcal{M}/\mu}$, which is given by:

$$\zeta^{Comp}_{\mathcal{M}/\mu} = \mathbb{E}_{pc}\left[\left(\frac{\bar{\delta}}{\bar{\epsilon}} - \frac{\mathcal{M}_\theta}{\mu_\theta}\right)\left(\chi^{FPT}_\theta - \chi_\theta\right)\right]\left(\mathbb{E}_{pc}\left[\frac{\mu_\theta - \mathcal{M}_\theta}{\mu_\theta \mathcal{M}_{o,e}}\right] - \mathbb{E}_{pc}\left[\frac{1}{\mu_\theta}\right]\frac{\bar{\epsilon} - \bar{\delta}}{\bar{\delta}}\right) \tag{42}$$

It is easy to verify that a sufficient condition for $\chi^{FPT}_\theta - \chi_\theta$ to be positive for all firms θ is that $\frac{\mathcal{M}_\theta}{\mu_\theta}$ is monotonous in productivity, and firms with higher market power simultaneously have lower pass-through rates.[24] In this case, the presence of monopsonistic competition magnifies the pro-competitive effects of trade highlighted in the monopolistic competition literature. In the converse case, market expansion may have anti-competitive effects, even if markups and price pass-through rates are negatively correlated. In other words, well-established conditions for the existence of pro-competitive effects of trade seize to be sufficient when firms have heterogeneous degrees of labor market power.

The second term is labeled $\zeta^{Demand}_{\mathcal{M}/\mu}$ and pertains to the previous discussion, which showed that firms with lower wage pass-through and more labor market power tend to raise markdowns in response to an increase in aggregate demand.

$$\zeta^{Demand}_{\mathcal{M}/\mu} = \mathbb{E}_{pc}\left[\left(1 - \frac{\bar{\epsilon}}{\bar{\delta}}\frac{\mathcal{M}_\theta}{\mu_\theta}\right)\left(\chi^{FPT}_\theta - \chi_\theta \frac{1}{\gamma_\theta}\right)\frac{\bar{\epsilon} - \bar{\delta}}{\beta_\theta}\right] \tag{43}$$

First, note that this effect only operates under imperfect competition in labor markets and vanishes as $\beta_\theta \to \infty$. In general $\zeta^{Demand}_{\mathcal{M}/\mu}$ tends to have opposing effects to $\zeta^{Comp}_{\mathcal{M}/\mu}$. Intuitively, the exact firms that respond to rising competition by raising their markdowns also tend to increase markdowns as aggregate demand rises.

The following corollary summarizes a set of sufficient conditions for changes in markups to yield welfare-enhancing reallocations.

Corollary 2. *$\zeta_{\mathcal{M}/\mu} > 0$ requires firms with higher profit-margins $\frac{\mu_\theta}{\mathcal{M}_\theta}$ to have lower wage and price pass-throughs, wage pass-throughs to satisfy $\mathbb{E}_{pc}[\chi^{FPT}_\theta - \chi_\theta \frac{1}{\gamma_\theta}] > 0$, entry to be initially excessive, and labor supply to be initially excessive.*

Changes in entry ζ_{θ^*} Sufficient conditions for market expansion to yield beneficial changes in selection were already provided in Proposition 4. When preferences are aligned, then competition forces welfare-enhancing changes in selection.

[24] For markdowns recent evidence provided by, e.g., Berger *et al.* (2022a),Hershbein *et al.* (2022) and Seegmiller (2021) suggests this is indeed the case. For markups, this is documented by, e.g., Amiti *et al.* (2019).

Corollary 3. *Sufficient conditions for market expansion to cause welfare enhancing changes in selection are as in Proposition 4.*

4.3.3 Discussion

When the conditions stated in corollaries 1, 2, and 3 are jointly satisfied, then market expansion reduces aggregate distortions and provides gains that are larger than in an economy with homogeneous markdowns and markups. In this case, Table 4.1 provides a lower bound for the real income gains from market expansion.

It is worth discussing how the stated conditions relate to empirically documented patterns of markdowns, markups, and pass-throughs across firms.

1. $\frac{\mathcal{M}_\theta}{\mu_\theta}$: Together, corollaries 1 to 3 require markdown to markup ratios to be increasing in firm productivity. Note that this does not require markups and markdowns, respectively, to be strictly increasing in firm size. Tortarolo & Zarate (2018), Brooks *et al.* (2021) and Hershbein *et al.* (2022) extend the "production" approach to estimating markups[25] to estimate plant-level markdowns and markups in Columbia, India and China, and the US respectively. Broadly speaking, these papers find that markups and markdowns are correlated with each other across firms, and respectively positively related to firm size and firm productivity.[26] While far from conclusive, this evidence suggests that the theoretical condition is empirically plausible.

2. Pass-throughs $\rho_\theta, \gamma_\theta$: For market expansion to cause welfare-enhancing reallocations, the model requires wage and price pass-throughs to be decreasing in markdowns and markups. The exchange rate pass-through literature (Berman *et al.* (2012), Amiti *et al.* (2019)) documents that firms with greater sales share have lower price pass-through. Berger *et al.* (2022a) provide evidence that wage pass-throughs are decreasing in firms' payroll shares. Together, these findings suggest that the theoretical conditions highlighted above are empirically plausible.[27]

3. Rents $\epsilon_\theta, \delta_\omega$: Whether or not preferences are aligned has not been empirically documented one way or another.[28]

[25] Introduced by Loecker (2011), this approach combines insights from Hall (1988) with production function estimation techniques from the IO literature (Levinsohn & Petrin (2003), Loecker & Warzynski (2012), Ackerberg *et al.* (2015)).

[26] To be more precise, Hershbein *et al.* (2022) document an inverse U-shaped relationship.

[27] It is worth noting that in all the papers mentioned, sales and wage bills are sufficient proxies for firm productivity by assumption.

[28] Recent work by Lamadon *et al.* (2022) and Kroft *et al.* (2020) estimates the size of rents in labor markets. However, the structural models used in these papers are equivalent to assuming that firm-level product demand and labor supply are isoelastic, implying that average worker rents arising from job ω equal $\frac{1}{1+\beta_\omega} w_\omega$.

Together, these considerations suggest that the conditions highlighted by the theoretical analysis have some empirical support. While a full quantitative analysis is beyond the scope of this paper, this suggests that policies promoting aggregate competition are effective tools to counter distortions from imperfect competition. Further, the gains from market expansion are likely larger than existing estimates suggest. In this context, an evaluation of the gains from trade in a model that explicitly accounts for trade costs might be of independent interest.

Lastly, Proposition 5 can be used directly to calculate counterfactual changes in real income via appropriate aggregates of firm-level elasticities. In the appendix, I discuss how reduced-form cross-sectional estimates of pass-throughs in wages and prices can be used to inform these elasticities.

5 Extensions

This section discusses extensions of the baseline framework along multiple dimensions. Details are relegated to Appendix C. First, I discuss 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 alternative labor supply systems that also feature variable labor supply elasticities.

Local Labor Markets The model presented in the main text models an economy with an integrated goods and a fully integrated labor market. Thereby, the model abstracts from local labor markets. Appendix C.1 presents an extension of the model that features segmented regional labor markets. Firms sell goods in the national output market, but hire workers in the local labor markets. Entry and selection are endogenous in each labor market. I show that in this environment, under inelastic labor supply, the market allocation is socially optimal so long as markdowns and markups are homogeneous across all the economy. The extension shows that the model ingredients are well suited to study the heterogeneous impacts of economy-wide labor market policies, such as a common minimum wage, on local labor market outcomes: The mass of employers, the selection cutoff, and differences in employment concentration across employers.

Heterogeneous worker groups Appendix C.2 extends the model to allows for heterogeneous worker groups, for example heterogeneity in skill or occupations, and derives analogous conditions under which the resulting market equilibrium is efficient.

VES type labor supply Appendix C.3 provides a model with VES-type labor supply. I show that the efficiency of the market allocation remains tied to homogeneous markdowns

and markups.

6 Conclusion

A growing body of emprical evidence highlights the prevalence of monopsony in labor markets, raising concerns about its implications for welfare and inequality. As a growing body of evidence finds that labor market power is not only sizeable but also varies substantially across firms, the welfare consequences of monopsony in the presence of firm heterogeneity have become a first-order concern for academics and policy-makers alike. This paper proposes a new framework suited to analyze misallocations caused by imperfect competition in labor markets in the presence of firm (and worker) heterogeneity.

The paper shows that monopsony introduces no allocative distortions when labor market power (i.e., wage markdowns) is constant across firms. In this case, distortions are common across firms and, thus, lead to no misalignment of social and private benefits to production of a variety. General equilibrium forces ensure that the externalities introduced by imperfect competition in the labor market are internalized, and the market optimally selects the number of entrants, the selection cutoff for exit, and the relative labor allocation across firms. Conversely, when labor market power varies across firms, distortions are no longer equalized across pr oducers, and the market cannot internalize the relevant externalities.

An emerging set of stylized facts suggests labor market power varies substantially across firms. The theoretical results, thus, suggest that industrial policy has ample opportunities to improve welfare by reallocating resources across producers. The paper analysis two such policies - taxation and market integration. The model sheds light on how micro-level heterogeneity in markdowns and markups shapes the effect of these policies. Overall, I find that increasing competition through market expansion indeed has the potential to raise welfare and reduce allocative distortions. Firm-level taxation, on the other hand, is likely hard to implement.

The contributions of this paper provide exciting avenues for future research. First, the model is well suited to help quantify the impact of monopsony on welfare, entry, and the gains from trade. The model also provides a starting point to analyze the effectiveness of competing policies proposals meant to address a lack of competition in labor markets - e.g., taxes on wage income or profits, entry subsidies, antitrust or minimum wages - regarding their impact on welfare and wage inequality.

References

Ackerberg, Daniel A., Kevin Caves and Garth Frazer, 'Identification Properties of Recent Production Function Estimators.' *Econometrica*, **83**, pp. 2411–2451, *2015*.

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

Anderson, Simon, Andre de Palma and Jacques Thisse, 'The CES and the logit: Two related models of heterogeneity.' *Regional Science and Urban Economics*, **18** (1), pp. 155–164, *1988*.

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*.

Ashenfelter, Orley, David Card, Henry S. Farber and Michael R. Ransom, 'Monopsony in the Labor Market: New Empirical Results and New Public Policies.' working paper, *2021*.

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

Bachmann, Ronald, Gőkay Demir and Hanna Frings, 'Labour Market Polarisation, Job Tasks and Monopsony Power.' IZA Discussion Papers 13989, Institute of Labor Economics (IZA), *2020*.

Baqaee, David, Emmanuel Fahri, and Kunal Sangani, 'The Darwinian Returns to Scale.' working paper, *2021*.

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

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

_ , _ and _ , 'Minimum Wages, Efficiency and Welfare.' NBER Working Papers 29662, National Bureau of Economic Research, Inc, *2022b*.

Berman, Nicolas, Philippe Martin and Thierry Mayer, ' How do Different Exporters React to Exchange Rate Changes? *.' *The Quarterly Journal of Economics*, **127** (1), pp. 437–492, *2012*, DOI: http://dx.doi.org/10.1093/qje/qjr057.

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.

Burdett, Kenneth and Dale T Mortensen, 'Wage Differentials, Employer Size, and Unemployment.' *International Economic Review*, **39** (2), pp. 257–273, *1998*.

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*.

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*.

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, *2018*.

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*.

Ghironi, Fabio and Marc J. Melitz, 'International Trade and Macroeconomic Dynamics with Heterogeneous Firms.' *The Quarterly Journal of Economics*, **120** (3), pp. 865–915, *2005*.

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

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*.

Helpman, Elhanan and Paul Krugman, *Market Structure and Foreign Trade: Increasing Returns, Imperfect Competition, and the International Economy*, **1** of MIT Press Books: The MIT Press, *1987*.

Hershbein, Brad, Claudia Macaluso and Chen Yeh, 'Monopsony in the U.S. labor market.' working paper, *2022*.

Jarosch, Gregor, Jan Sebastian Nimczik and Isaac Sorkin, 'Granular Search, Market Structure, and Wages.' IZA Discussion Papers 12574, Institute of Labor Economics (IZA), *2019*.

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.

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.

Levinsohn, James and Amil Petrin, 'Estimating Production Functions Using Inputs to Control for Unobservables.' *Review of Economic Studies*, **70** (2), pp. 317–341, *2003*.

Loecker, Jan De, 'Product Differentiation, Multiproduct Firms, and Estimating the Impact of Trade Liberalization on Productivity.' *Econometrica*, **79** (5), pp. 1407–1451, *2011*, DOI: http://dx.doi.org/ECTA7617.

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

MacKenzie, Gaelan, 'Trade and Market Power in Product and Labor Markets.' working paper, *2018*.

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*.

_ , 'Imperfect Competition in the Labor Market.' **4B**: Elsevier, 1st edition, Chapter 11, pp. 973–1041, *2011*.

_ , '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*.

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*.

Seegmiller, Bryan, 'Valuing Labor Market Power: The Role of Productivity Advantages.' working paper, *2021*.

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 Ushev, 'When can a demand system, be described by a multinomial logit with income effect?' working paper, *2016*.

Tortarolo, Dario and Roman David Zarate, 'Measuring Imperfect Competition in Product and Labor Markets. An Empirical Analysis using Firm-level Production Data.' working paper, *2018*.

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*.

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.

A Derivations

A.1 Problem of the Household

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

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

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:

$$U_c C = -\lambda_C \int \Upsilon'_\theta(\frac{c_\theta}{C})\frac{c_\theta}{C}dM^C(\theta) \tag{44}$$

$$U_N N = \lambda_N \int \Psi'_\omega(\frac{n_\omega}{N})\frac{n_\omega}{N}dM^E(\omega) \tag{45}$$

$$-\lambda_C \Upsilon'_\theta(\frac{c_\theta}{C})\frac{1}{C} = \gamma p_\theta \tag{46}$$

$$\lambda_N \Psi'_\omega(\frac{n_\omega}{N})\frac{1}{N} = \gamma w_{\omega'} \tag{47}$$

Using (44) to substitute λ_C in (46) 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:

$$\frac{p_\theta}{\mathcal{P}} = \Upsilon'_\theta(\frac{c_\theta}{C}). \tag{48}$$

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:

$$\frac{w_\omega}{\mathcal{W}} = \Psi'_\omega(\frac{n_\omega}{N}). \tag{49}$$

A.2 Microfoundation of the Labor Supply System

Kimball labor supply There is a continuum of workers i of mass L. Workers optimally pick one firm ω. Supposing that workers' preferences of hours worked and consumption are separable, I analyze the employer choice problem assuming 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:

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

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 Formulation of the Overhead Goods Sector

Instead of assuming a separate sector producing overhead goods, the model can be formulated by allowing for job differentiation within firms. Suppose that we express the labor supply index N as:

$$1 = M\Psi_e(\frac{n}{N}) + M\int_{\theta^*}^{\infty}(\Psi_\theta(\frac{n_\theta}{N}) + \Psi_{\theta,o}(\frac{n_{\theta,o}}{N}))dG(\theta), \tag{50}$$

where n_θ denotes the jobs in variable production at firm θ, while $n_{\theta,o}$ denotes jobs in the production of overhead goods. 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. To keep the firm problem equally tractable, assume that workers have to work in the jobs that they were hired for. That is a worker hired for overhead cannot be used in variable 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: $w_{\theta,o} = \mathcal{W}\Psi'_{\theta,o}(\frac{f_o}{A_{\theta,o}LN})$, where $A_{\theta,o}$ is a productivity shifter that allows firm type to influence the efficiency at which firms produce overhead. If $\Psi_{\theta,o}(x) = \Psi_o(x)$ and $A_{\theta,o} = A_o$, then this formulation is isomorphic to the one presented in the main text. If, instead, $\Psi_{\theta,o}(x)$ differs across final good producers, then this formulation is in the cross-section isomorphic to a model where overhead costs vary across firms. In this case, firms decide to produce if, and only, if:

$$X_\theta = \frac{(1-\frac{\mathcal{M}_\theta}{\mu_\theta})p_\theta c_\theta}{w_{\theta,o}} \geq f_o/L. \tag{51}$$

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

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

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. The same firm-level elasticities determine the margins of inefficiency when markups and markdowns vary across firms, but account for the fact that inframarginal employment surpluses from overhead now differ across employers. The notion of aligned preferences in Section 4 remains conceptually unchanged, once Assumption 1 is imposed.

Finally, the model can be formulated by doing away the entry sector as well. To this end, assume the same specification of preferences as in 50. Prospective sellers of final goods have to hire workers to produce entry goods in quantity f_e using the same linear technology in labor.

B Proofs

B.1 Proof of Theorem 1

Proof. The proof assumes that an equilibrium exists and is unique. Conditions ensuring equilibrium existence and uniqueness are provided in Proposition 3.

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 $\forall\theta \in \text{support}\{G(\theta)\}, \epsilon_\theta \equiv \mu = \frac{\sigma}{\sigma-1}$ and $\forall\omega' \in \{o,e,\{\theta\}_{\theta\in\text{support}\{G(\theta)\}}\}, \delta_{\omega'} \equiv \mathcal{M} = \frac{\beta}{\beta+1}$.

First, I state the planner's problem in a more compact form. 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 per-capita $n_o = \frac{f_o}{L}$ and $n_e = \frac{f_e}{L}$ to the production of entry and overhead goods for each entering and producing variety. Thee problem of the planner can be stated as:

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

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:

$$\Upsilon(\frac{c_\theta}{C}) = \frac{\epsilon_\theta}{\delta_\theta}(\frac{\lambda_N}{\lambda_C})\Psi(\frac{c_\theta}{\bar{N}A_\theta}). \tag{52}$$

The first order condition with respect to M_e implies:

$$\frac{\lambda_N}{\lambda_C} = \frac{\int_{\theta^*}^{\infty}\Upsilon(\frac{c_\theta}{C})dG(\theta)}{\Psi(f_e/(L\bar{N})) + \int_{\theta^*}^{\infty}\left\{\Psi(f_o/(L\bar{N})) + \Psi(\frac{n_\theta}{\bar{N}})\right\}dG(\theta)} = 1, \tag{53}$$

where the last equality follows imposing that all constraints bind.

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

$$U_C C = -\lambda_c M\int_{\theta^*}^{\infty}\Upsilon'(\frac{c_\theta}{C})\frac{c_\theta}{C}dG(\theta) \tag{54}$$

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

$$\lambda_C\Upsilon(\frac{c_{\theta^*}}{C}) - \lambda_N\Psi(\frac{c_{\theta^*}}{\bar{N}A_{\theta^*}}) = \lambda_N\Psi\left(\frac{f_o}{NL}\right) \tag{55}$$

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:

$$p_\theta = \frac{\mu_\theta}{\mathcal{M}_\theta}\frac{w_\theta}{A_\theta}. \tag{56}$$

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 56 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 (53) into (52) shows that this also coincides with the planner's first-order condition pinning down relative firm sizes (52). Thus, conditional on C, the planner and the market choose the same relative firm-level allocations across consumption good producers.

Next, I use (53) to define the "entry" condition of the planner:

$$\int_{\theta^*}^{\infty}(\Upsilon(\frac{c_\theta}{C}) - \Psi(\frac{n_\theta}{N}) - \Psi(f_o/(L\bar{N})))dG(\theta) = \Psi(f_e/(LN)), \tag{57}$$

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

$$\int_{\theta^*}^{\infty} \left(L \left(\mathcal{P} \Upsilon'_\theta(\frac{c_\theta}{C}) - \mathcal{W} \frac{1}{A_\theta} \Psi'_\theta(\frac{c_\theta}{\bar{N} A_\theta}) \right) c_\theta - \mathcal{W} \Psi'_o(\frac{f_o}{L\bar{N}}) f_o \right) dG(\theta) = \mathcal{W} \Psi'_e(\frac{f_e}{L\bar{N}}) f_e. \tag{58}$$

Under constant markups and markdowns, (57) and (58) coincide. To see this, divide (58) 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, $\left(\mathcal{P} \Upsilon'_\theta(\frac{c_\theta}{C}) - \mathcal{W} \frac{1}{A_\theta} \Psi'_\theta(\frac{c_\theta}{\bar{N} A_\theta}) \right) c_\theta = \Upsilon(\frac{c_\theta}{C}) - \Psi(\frac{n_\theta}{N})$, $\mathcal{W} \Psi'_o(f_o/(LN)) f_o/L = \Psi(\frac{f_o}{NL})$ and $\mathcal{W} \Psi'_e(f_e/(LN)) f_e/L = \Psi(\frac{f_e}{NL})$. Thus, (57) and (58) provide the same restriction on entry. Analogous derivations imply that the planner's FOC pinning down θ^* (55) is equivalent to the market's selection equation in (27).

This establishes the if part of the theorem: Free entry then 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.[29] By the previous arguments, this establishes that the planner and the market allocation coincide.

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

$$\Upsilon(\frac{c_\theta}{C}) = \frac{\mu_\theta}{\mathcal{M}_\theta} \Psi(\frac{c_\theta / A_\theta}{N}) \frac{\mathbb{E}_{pc\epsilon}\left[\frac{1}{\epsilon_\theta}\right] \epsilon_\theta}{\mathbb{E}_{wn\delta}\left[\frac{1}{\delta}\right] \delta_\theta}. \tag{59}$$

Comparing equations (52) and (59), it is evident that a necessary condition for market and planner allocation to coincide is that $\frac{\mu_\theta}{\mathcal{M}_\theta} \frac{\mathbb{E}_{pc\epsilon}\left[\frac{1}{\epsilon_\theta}\right]}{\mathbb{E}_{wn\delta}\left[\frac{1}{\delta}\right]} = 1$. □

B.2 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

[29] 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 $U_C C \mu = -\lambda_c$.

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, $\frac{\mu_\theta}{\mathcal{M}_\theta} > \frac{\mu_{\theta'}}{\mathcal{M}_{\theta'}}$. □

B.3 Proof of Lemma 2

Proof. Denoting $d\log c_\theta$ the change in per-capita quantity consumed from the reallocation, note 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 relative firm sizes do not change by design , $d\log c_\theta = d\log\tilde{c}$. Since we reduce output per-variety, $-d\log c_\theta = -d\log\tilde{c} > 0$. □

B.4 Proof of Lemma 3

Proof. Suppose the selection cutoff increases by $d\theta^* > 0$. Using Equation (70), $d\log NL = 0$ and that reallocation proportionally increases the size of each remaining firm, we obtain that this reallocation raises welfare, $d\log C > 0$, if, and only if,

$$-\omega^{pc}_{\theta^*}\left(\epsilon_{\theta^*} - \bar{\epsilon} - \frac{\mathcal{M}_{\theta^*}}{\mu_{\theta^*}}(\delta_{\theta^*} - \bar{\delta}) - (1 - \frac{\mathcal{M}_\theta}{\mu_{\theta^*}})(\delta_o - \bar{\delta})\right)\frac{g(\theta^*)}{1 - G(\theta^*)}d\theta^* < 0, \tag{60}$$

which provides the inequality used in the main text. □

B.5 Proof of Theorem 7

Proof. Following the arguments in Appendix B.1, it follows that conditional on C and N the planner's and the market allocation choose the same relative sizes of firms, number of entrants, and selection. The first order condition of the planner with respect to N is given by $U_N N = \lambda_N M_e \int_\omega \Psi'(\frac{n_\omega}{N})\frac{n_\omega}{N}d\omega = \lambda_N\frac{1}{\mathcal{M}}$, which combining with the first order condition for C and N implies that the planner chooses C and N according to the following F.O.C.:

$$-\frac{U_C C}{U_N N} = \frac{\mathcal{M}}{\mu}. \tag{61}$$

The market, in turn, sets $-\frac{U_C C}{U_N N} = 1$. The proposed leisure tax restores the equivalence between the planner's choice and the household's labor-leisure choice, ensuring that

market and socially optimal allocations coincide. □

B.6 Proof of Lemma 4

Proof. We raise aggregate labor supply, allocating the additional labor in a manner that ensures that the reduction in entry cost does not mechanically raise income by causing aggregate profits to become positive. This, in turn, can be ensured by setting $d\log c_\theta = d\log \tilde{c}$ such that:

$$-\mathbb{E}_{pc}[1-\frac{\bar{\epsilon}}{\bar{\delta}}\frac{\mathcal{M}_\theta}{\mu}]d\log\tilde{c} = \mathbb{E}_{pc}[1-\frac{\mathcal{M}_\theta}{\mu_\theta}]\frac{\mathbb{E}_{pc}\left[\frac{1}{\mu_\theta}\right]^{-1}}{\mathcal{M}_{o,e}}d\log N$$

The reallocation, then raises welfare if $d\log\frac{C}{N} = \left[\frac{\bar{\epsilon}-\bar{\delta}}{\bar{\delta}} - \mathbb{E}_{pc}[1-\frac{\mathcal{M}_\theta}{\mu_\theta}]\frac{\mathbb{E}_{pc}\left[\frac{1}{\mu_\theta}\right]^{-1}}{\mathcal{M}_{o,e}}\right]d\log N > 0$,

which requires $\frac{\bar{\epsilon}-\bar{\delta}}{\bar{\delta}} - \mathbb{E}_{pc}[1-\frac{\mathcal{M}_\theta}{\mu_\theta}]\frac{\mathbb{E}_{pc}\left[\frac{1}{\mu_\theta}\right]^{-1}}{\mathcal{M}_{oe}} > 0$, as stated in the main text. □

B.7 Market Size and welfare.

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.

B.7.1 Setting up the system of equations

Differentiating the consumtion and labor indices, we obtain:

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

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

Differentating the wage and price aggregates:

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

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

where $\frac{1}{\tilde{\mathcal{M}}_f} \equiv \frac{(1-G(\theta^*))p_o f_o}{(1-G(\theta^*))p_o f_o + p_e f_e}\frac{1}{\mathcal{M}_o} + \frac{p_e f_e}{(1-G(\theta^*))p_o f_o + p_e f_e}\frac{1}{\mathcal{M}_e}$ 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:

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

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 cutoff condition:

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

B.7.2 Solving the system

First, we 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:

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

$$\Leftrightarrow d\log \frac{c_\theta}{C} = \sigma_\theta d\log \mathcal{P} - \sigma_\theta \left(\frac{1}{\sigma_\theta}\frac{1-\rho_\theta}{\rho_\theta} d\log(\frac{c_\theta}{C}) + \frac{1}{\gamma_\theta \beta_\theta} d\log \frac{c_\theta}{C} + \frac{1}{\gamma_\theta \beta_\theta} d\log \frac{C}{N} + d\log \mathcal{W} \right),$$

where the last term utilized the fact that $d\log n_\theta = d\log c_\theta$. Collecting the relevant terms shows that changes in relative firm quantitities are entirely determined by aggregates, but depend on the invidiual firm elasticities:

$$d\log(\frac{c_\theta}{C}) = -\frac{\sigma_\theta \beta_\theta \rho_\theta \gamma_\theta}{\gamma_\theta \beta_\theta + \rho_\theta \sigma_\theta} d\log \mathcal{W}/\mathcal{P} - \frac{\sigma_\theta \rho_\theta}{\gamma_\theta \beta_\theta + \rho_\theta \sigma_\theta} d\log C/N \equiv -\chi_\theta d\log \mathcal{W}/\mathcal{P} - \frac{\chi_\theta}{\beta_\theta \gamma_\theta} d\log C/N \tag{62}$$

We can also use this to derive changes in firms profitability:

$$d\log \frac{\mu_\theta}{\mathcal{M}_\theta} = -\left(1 - \rho_\theta \gamma_\theta \frac{\beta_\theta + \sigma_\theta}{\sigma_\theta \rho_\theta + \gamma_\theta \beta_\theta}\right) d\log \mathcal{W}/\mathcal{P} - \frac{1}{\beta_\theta \gamma_\theta}\left(\gamma_\theta - \rho_\theta \gamma_\theta \frac{\beta_\theta + \sigma_\theta}{\sigma_\theta \rho_\theta + \gamma_\theta \beta_\theta}\right) d\log(\frac{C}{N}) \tag{63}$$

Relative price indices: Subtracting the price indices yields:

$$\mathbb{E}_{pc}\left[1 - \frac{\mathcal{M}_\theta}{\mu_\theta}\right] \frac{1}{\tilde{\mathcal{M}}_f} d\log NL = d\log \mathcal{W}/\mathcal{P} - \mathbb{E}_{pc}\left[\frac{1}{\sigma_\theta}\right] d\log C/N, \tag{64}$$

Free entry condition We substitute the expression for profits to obtain:

$$\begin{aligned} &\mathbb{E}_{pc}\left[\left(1 - \tfrac{\mathcal{M}_\theta}{\mu_\theta}\right)\tfrac{1}{\mathcal{M}_\theta} d\log\left(\tfrac{c_\theta}{C}\right)\right] + \mathbb{E}_{pc}\left[\left(1 - \tfrac{\mathcal{M}_\theta}{\mu_\theta} + \tfrac{1}{\mu_\theta}\right) d\log \tfrac{\mu_\theta}{\mathcal{M}_\theta}\right] \\ &= -\mathbb{E}_{pc}\left[1 - \tfrac{\mathcal{M}_\theta}{\mu_\theta}\right] \tfrac{1}{\mathcal{M}_o} d\log(NL) - \mathbb{E}_{pc}\left[\left(1 - \tfrac{\mathcal{M}_\theta}{\mu_\theta}\right)\tfrac{1}{\mathcal{M}_\theta}\right] d\log \tfrac{C}{N} \end{aligned} \tag{65}$$

Consumption and labor index: The consumption index is already expressed only in terms of aggregates:

$$\mathbb{E}_{pc}\left[\epsilon_\theta\right] d\log M - \omega^{pc}_{\theta^*}\epsilon_{\theta^*}\frac{g(\theta^*)}{1-G(\theta^*)}d\theta^* + \mathbb{E}_{pc}\left[d\log(\frac{c_\theta}{C})\right] = 0 \tag{66}$$

For the labor index, we can write:

$$\mathbb{E}_{pc}\left[1-\frac{\mathcal{M}_\theta}{\mu}\right] d\log NL = \mathbb{E}_{wn}[\delta] d\log M + \mathbb{E}_{pc}\left[\frac{\mathcal{M}_\theta}{\mu_\theta} d\log(\frac{c_\theta}{C})\right] + \mathbb{E}_{pc}\left[\frac{\mathcal{M}_\theta}{\mu_\theta}\right] d\log\frac{C}{N} - \omega^{pc}_{\theta^*}\left((1-\frac{\mathcal{M}_{\theta^*}}{\mu_{\theta^*}})\delta_o + \frac{\mathcal{M}_{\theta^*}}{\mu_{\theta^*}}\delta_{\theta^*}\right)d\theta^*, \tag{67}$$

B.7.3 Proof of Proposition 3

In the CES economy, changes in $\frac{C}{N}$ equal changes in $\mathcal{W}/\mathcal{P}$, so we can use (64) to infer changes in welfare as a function of changes in the net labor supply. Next, we use the condition for household optimization, $U_C C = -U_N N$, to relate changes in C to changes in N : $d\log C = \phi d\log N$, where ϕ is defined as in the statement of the proposition. This allows solving for C and N as a function of $d\log L$ via equation (64), which in turn implies the expression in the main text expression stated in the text.

B.7.4 Proof of Proposition 4

Proof. The definition of the price and wage aggregate implies that we can write: $\frac{C}{N} = \frac{\mathbb{E}_{pc}[\epsilon_\theta]}{\mathbb{E}_{wn}[\delta]}\frac{\mathcal{W}}{\mathcal{P}}$. Changes in real income are, given by:

$$d\log\frac{C}{N} = d\log(\mathbb{E}_{pc}\left[\epsilon_\theta\right]/\mathbb{E}_{wn}\left[\delta\right]) + d\log\mathcal{W}/\mathcal{P}.$$

Changes in the total amount of households rents, in turn, equal: $d\log(\mathbb{E}_{pc}\left[\epsilon_\theta\right]/\mathbb{E}_{wn}\left[\delta\right]) = \mathbb{E}_{pc}\left[\frac{1}{\mu_\theta}\right] d\log\frac{C}{N} - \mathbb{E}_{pc}[1-\frac{\mathcal{M}_\theta}{\mu_\theta}]\frac{1}{\mathcal{M}_{o,e}}(d\log NL)$, implying: $\mathbb{E}_{pc}[\frac{1}{\sigma_\theta}] d\log\frac{C}{N} = d\log\mathcal{W}/\mathcal{P} - \mathbb{E}_{pc}[1-\frac{\mathcal{M}_\theta}{\mu_\theta}]\frac{1}{\mathcal{M}_{o,e}}(d\log NL)$. Aggregating changes in relative firm sizes using equation (62), we also have that"

$$d\log\frac{\mathcal{W}}{\mathcal{P}} + \mathbb{E}_{pc}[\frac{1}{\sigma_\theta}]\mathbb{E}_{pc}\left[d\log\frac{c_\theta}{C}\right] = \frac{\mathbb{E}_{pc}\left[\chi_\theta\frac{1}{\beta_\theta\gamma_\theta}\right]}{\mathbb{E}_{pc}\left[\chi_\theta(1+\frac{1}{\beta_\theta\gamma_\theta})\right]}\mathbb{E}_{pc}[1-\frac{\mathcal{M}_\theta}{\mu_\theta}]\frac{1}{\mathcal{M}_{o,e}}(d\log NL)$$

Ignoring changes in the selection cutoff for now, the derivatives of the utility and labor supply aggregates, in turn, imply that aggregate changes in relative firm sizes equal $\mathbb{E}_{pc}\left[d\log\frac{c_\theta}{C}\right] = \frac{\bar{\epsilon}}{\bar{\delta}}\mathbb{E}_{pc}\left[\frac{\mathcal{M}_\theta}{\mu_\theta} d\log\frac{n_\theta}{N}\right] - \frac{\bar{\epsilon}}{\bar{\delta}}\mathbb{E}_{pc}[1-\frac{\mathcal{M}_\theta}{\mu_\theta}](\frac{d\log NL}{d\log L})$. Substituting the change in relative

firm sizes, we obtain:

$$d\log\frac{\mathcal{W}}{\mathcal{P}}=\frac{\left[1+\frac{\mathbb{E}_{pc}\left[\chi_\theta\frac{1}{\beta_\theta\gamma_\theta}\right]}{\mathbb{E}_{pc}\left[\chi_\theta(1+\frac{1}{\beta_\theta\gamma_\theta})\right]}\frac{\bar\delta}{\bar\epsilon}\mathbb{E}_{pc}\left[\frac{\mathcal{M}_\theta}{\mu_\theta}\right]^{-1}+\frac{\mathbb{E}_{pc}[\frac{1}{\sigma_\theta}]}{\mathbb{E}_{pc}\left[\frac{\mathcal{M}_\theta}{\mu_\theta}\right]}\mathcal{M}_{o,e}-\frac{\mathbb{E}_{pc}\left[\frac{\mathcal{M}_\theta}{\mu_\theta}\chi_\theta\frac{1}{\beta_\theta\gamma_\theta}\right]}{\mathbb{E}_{pc}\left[\frac{\mathcal{M}_\theta}{\mu_\theta}\right]}\right]}{(1+\frac{\bar\delta}{\bar\epsilon}\mathbb{E}_{pc}\left[\frac{\mathcal{M}_\theta}{\mu_\theta}\right]^{-1}-\frac{\mathbb{E}_{pc}\left[\frac{\mathcal{M}_\theta}{\mu_\theta}\chi_\theta\right]}{\mathbb{E}_{pc}\left[\frac{\mathcal{M}_\theta}{\mu_\theta}\right]}-\frac{\mathbb{E}_{pc}\left[\frac{\mathcal{M}_\theta}{\mu_\theta}\chi_\theta\frac{1}{\beta_\theta\gamma_\theta}\right]}{\mathbb{E}_{pc}\left[\frac{\mathcal{M}_\theta}{\mu_\theta}\right]})}\mathbb{E}_{pc}[1-\frac{\mathcal{M}_\theta}{\mu_\theta}]\frac{1}{\mathcal{M}_{o,e}}(\frac{d\log NL}{d\log L})$$

Gains from market expansion are ensured if the following inequality holds:

$$-\frac{\mathbb{E}_{pc}\left[\chi_\theta\right]}{\mathbb{E}_{pc}\left[\chi_\theta(1+\frac{1}{\beta_\theta\gamma_\theta})\right]}\frac{\bar\delta}{\bar\epsilon}+\mathbb{E}_{pc}[\frac{\mathcal{M}_{o,e}}{\sigma_\theta}]>-\mathbb{E}_{pc}\left[(\sigma_\theta-1)\frac{\mathcal{M}_\theta}{\sigma_\theta}\chi_\theta\right]\tag{68}$$

I now establish that this inequality holds for any θ. First, as $\beta_\theta\to 0$, this equality this inequality always holds, so long as $\frac{\mathcal{M}_{o,e}}{\sigma_\theta}>0$. Conversely, as $\beta_\theta\to\infty$, $-\frac{1}{\bar\epsilon}-\frac{1}{\mu_\theta}>-\sigma_\theta$ requiring that $\sigma_\theta>2$. Monotonicity implies establishes that (68) holds individually for θ, for any labor supply elasticity. Similarly, the inequality holds individually for any σ_θ : For $\sigma_\theta\to\infty$, this requires $-\frac{1}{1+\frac{1}{\beta_\theta\gamma_\theta}}\bar\delta>-\infty$, while for $\sigma_\theta\to 0$ it is also always satisfied. Since the inequality holds at any θ, Jensen's inequality can be applied to show that (68) holds.

To finalize the proof, I show under aligned preferences changes in the selection cutoff are always beneficial. Proposition 6 highlights that welfare effects of selection are captured by:

$$\zeta_{\text{selection}}=\frac{1}{\mathcal{M}_{o,e}}\iota_{\theta^*}\zeta_{\theta^*}\left(1-\mathbb{E}_{pc}\left[\frac{\mu_\theta-\mathcal{M}_\theta}{\mu_\theta}\right]\frac{\mu_{\theta^*}}{\mu_{\theta^*}-\mathcal{M}_{\theta^*}}\right)\frac{\bar\epsilon-\bar\delta}{\bar\delta}\mathbb{E}_{pc}\left[\frac{1}{\mu_\theta}\right].$$

where $\iota_{\theta^*}=\omega^{pc}_{\theta^*}\left(\epsilon_{\theta^*}-\bar\epsilon-\frac{\mathcal{M}_{\theta^*}}{\mu_{\theta^*}}\left(\delta_{\theta^*}-\bar\delta\right)-\left(1-\frac{\mathcal{M}_\theta}{\mu_{\theta^*}}\right)\left(\delta_o-\bar\delta\right)\right)$. Aligned preferences ensure that $\zeta_{\text{selection}}$ is always positive. If entrants have higher profit margins than the average firm, then $\frac{1}{\mathbb{E}_{pc}\left[\frac{1-\mathcal{M}_\theta}{\mu_\theta}\right]}-\frac{\mu_{\theta^*}}{\mu_{\theta^*}-\mathcal{M}_{\theta^*}}>0$. In this case, aligned preferences and the stated conditions imply that entrants also provide higher rents, ensuring that $\zeta_{\text{selection}}>0$. Conversely, if entrants have lower profit margins, then $1-\mathbb{E}_{pc}\left[\frac{\mu_\theta-\mathcal{M}_\theta}{\mu_\theta}\right]\frac{\mu_{\theta^*}}{\mu_{\theta^*}-\mathcal{M}_{\theta^*}}<0$, and $\iota_{\theta^*}<0$, again ensuring that any change in selection is beneficial to firms. □

B.7.5 Proof of Proposition 5

Proof. We substitute for firm level outcomes $d\log\frac{c_\theta}{C}$ in all relevant equilibrium equations to obtain two equations allowing us to pin down $d\theta^*$, $d\log\frac{C}{N}$ and $d\log\frac{W}{P}$ as a function of $d\log NL$. We first use (64) jointly with (62) and (63) to show that

$$d\log\frac{c_\theta}{C}=-\chi_\theta\alpha_\theta d\log\frac{C}{N}-\chi_\theta\mathbb{E}_{pc}\left[1-\frac{\mathcal{M}_\theta}{\mu_\theta}\right]\frac{1}{\mathcal{M}_f}d\log NL$$

$$d\log\frac{\mu_\theta}{\mathcal{M}_\theta}=-(1-\rho_\theta\gamma_\theta\frac{\beta_\theta+\sigma_\theta}{\sigma_\theta\rho_\theta+\gamma_\theta\beta_\theta})\mathbb{E}_{pc}\left[1-\frac{\mathcal{M}_\theta}{\mu_\theta}\right]\frac{1}{\mathcal{M}_f}d\log NL-\alpha_\theta\left(\gamma_\theta-\rho_\theta\gamma_\theta\frac{\beta_\theta+\sigma_\theta}{\sigma_\theta\rho_\theta+\gamma_\theta\beta_\theta}\right)d\log\frac{C}{N},$$

where $\alpha_\theta = \left[\frac{1}{\beta_\theta \gamma_\theta} + \mathbb{E}_{pc}\left[\frac{1}{\sigma_\theta}\right]\right]$ and $\chi_\theta = \frac{\sigma_\theta \beta_\theta \gamma_\theta \rho_\theta}{\rho_\theta \sigma_\theta + \beta_\theta \gamma_\theta}$.

Combining the selection with the entry condition, we obtain that changes in selection are proportional to changes in relative profits of the marginal and the average firm:

$$\mathbb{E}_\pi\left[d\log\left(\frac{\pi_\theta}{\pi_{\theta^*}}\right)\right] = \frac{1}{\zeta_{\theta^*}}\frac{g(\theta)}{1-G(\theta^*)}d\theta^*$$

Next, we apply the logic in Baqaee *et al.* (2021) to observe that $d\log\frac{\mathcal{W}}{P} - d\log\frac{C}{N}$ can be written as:

$$d\log\frac{\mathcal{W}}{\mathcal{P}} - d\log\frac{C}{N} = \underbrace{\mathbb{E}_{pc}\left[\frac{1-\mathcal{M}_\theta}{\mu_\theta}\right]\frac{1}{\mathcal{M}_f}d\log NL}_{\Delta\text{Wages for overhead and entry}} + \underbrace{\mathbb{E}_{pc}\left[\frac{1}{\mu_\theta}d\log\frac{c_\theta}{C}\right] - \mathbb{E}_{wn}\left[\frac{1}{\mathcal{M}_\theta}d\log\frac{n_\theta}{N}\right]}_{\Delta\text{average sales less wage cost for final goods}}$$

It follows that:

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

Substituting for $d\log\frac{\mathcal{W}}{\mathcal{P}}$ using Equation (64), we find:

$$\frac{g(\theta)}{1-G(\theta^*)}d\theta^* = -b_{\theta^*}\left(\frac{\mu_{\theta^*}}{\mu_{\theta^*}-\mathcal{M}_{\theta^*}} + \frac{1}{\mathbb{E}_{pc}\left[\frac{1-\mathcal{M}_\theta}{\mu_\theta}\right]}\right)\mathbb{E}_{pc}\left[\frac{1-\mathcal{M}_\theta}{\mu_\theta}\right]\frac{1}{\mathcal{M}_f}d\log NL - \zeta_{\theta^*}\left(\frac{\mu_{\theta^*}}{\mu_{\theta^*}-\mathcal{M}_{\theta^*}} - \frac{1}{\mathbb{E}_{pc}\left[\frac{1-\mathcal{M}_\theta}{\mu_\theta}\right]}\right)\mathbb{E}_{pc}\left[\frac{1}{\mu_\theta}\right]d\log\left(\frac{C}{N}\right) \tag{69}$$

Note that when the economy markdowns and markups are homogeneous, $d\log\frac{W}{P} = d\log\frac{C}{N}$, Equation (64) implies that there would be no change in selection.

Combining the derivatives of the utility and consumption indices with the expressions for relative changes in firm size yields:

$$\begin{aligned} 0 = & -\mathbb{E}\left[\frac{\mathcal{M}_\theta}{\mu_\theta} - \left(\frac{\bar\delta}{\bar\epsilon} - \frac{\mathcal{M}_\theta}{\mu_\theta}\right)\chi_\theta\alpha_\theta\right]d\log\frac{C}{N} \\ & -\frac{\bar\delta}{\bar\epsilon}\omega^{pc}_{\theta^*}\left(\epsilon_{\theta^*} - \bar\epsilon - \frac{\mathcal{M}_{\theta^*}}{\mu_{\theta^*}}(\delta_{\theta^*} - \bar\delta) - (1 - \frac{\mathcal{M}_\theta}{\mu_{\theta^*}})(\delta_o - \bar\delta)\right)\frac{g(\theta^*)}{1-G(\theta^*)}d\theta^* \\ & +\mathbb{E}_{pc}\left[\mathcal{M}_f - \left(\frac{\bar\delta}{\bar\epsilon} - \frac{\mathcal{M}_\theta}{\mu_\theta}\right)\chi_\theta\right]\mathbb{E}_{pc}\left[1 - \frac{\mathcal{M}_\theta}{\mu_\theta}\right]\frac{1}{\mathcal{M}_f}d\log(NL) \end{aligned} \tag{70}$$

Together, equations (69) and (70) can be used to derive the change in $d\log\frac{C}{N}$ as a function of changes in the effective labor supply $d\log NL$. With some manipulation, these expression can be rearranged to yield the expression in the theorem.

$$d\log\frac{C}{N} = \frac{\bar\epsilon - \bar\delta}{\bar\delta}d\log NL + \frac{\zeta_{\text{het}} + \zeta_{comp} + \zeta_{selection}}{\chi}d\log NL \tag{71}$$

where $\chi \equiv \mathbb{E}\left[\frac{\mathcal{M}_\theta}{\mu_\theta} - \left(\frac{\bar\delta}{\bar\epsilon} - \frac{\mathcal{M}_\theta}{\mu_\theta}\right)\chi_\theta\alpha_\theta\right] + \iota_{\theta^*}b_{\theta^*}\left(\frac{1}{\mathbb{E}_{pc}\left[\frac{1-\mathcal{M}_\theta}{\mu_\theta}\right]} - \frac{\mu_{\theta^*}}{\mu_{\theta^*}-\mathcal{M}_{\theta^*}}\right)\mathbb{E}_{pc}\left[\frac{1}{\mu_\theta}\right]$, and $\iota_{\theta^*} \equiv \omega^{pc}_{\theta^*}\left(\epsilon_{\theta^*} - \bar\epsilon - \frac{\mathcal{M}_{\theta^*}}{\mu_{\theta^*}}(\delta_{\theta^*} - \bar\delta\right.$

Further,

$$\zeta_{\text{entry}} = COV_{pc}\left[\left(\frac{\bar{\epsilon}}{\bar{\delta}}\frac{\mathcal{M}_\theta}{\mu_\theta} - 1\right)\chi_\theta^{FPT}, \frac{1}{\mu_\theta} - \frac{\bar{\delta}}{\bar{\epsilon}}\mathcal{M}_\theta\right] + \frac{\bar{\delta}}{\bar{\epsilon}}\mathbb{E}_{pc}\left[\left(\frac{\bar{\epsilon}}{\bar{\delta}} - \frac{\mu_\theta}{\mathcal{M}_\theta}\right)\chi_\theta^{FPT}\right]\left[\frac{\mathbb{E}_{pc}\left[\mathcal{M}_\theta\right] - \mathcal{M}_{o,e}}{\mathbb{E}_{pc}\left[\frac{1}{\mu_\theta}\right]^{-1}}\right]$$

$$\zeta_{\mathcal{M}/\mu} = \frac{\bar{\delta}}{\bar{\epsilon}}\mathbb{E}_{pc}\left[\left(1 - \frac{\bar{\epsilon}}{\bar{\delta}}\frac{\mathcal{M}_\theta}{\mu_\theta}\right)\left(\chi_\theta^{FPT} - \chi_\theta\right)\right]\left(\mathbb{E}_{pc}\left[\frac{\mu_\theta - \mathcal{M}_\theta}{\mu_\theta \mathcal{M}_{o,e}}\right] - \mathbb{E}_{pc}\left[\frac{1}{\mu_\theta}\right]\frac{\bar{\epsilon} - \bar{\delta}}{\bar{\delta}}\right) + \mathbb{E}\left[\left(1 - \frac{\bar{\epsilon}}{\bar{\delta}}\frac{\mathcal{M}_\theta}{\mu_\theta}\right)\left(\chi_\theta^{FPT} - \chi_\theta\frac{1}{\gamma_\theta}\right)\frac{\bar{\epsilon} - \bar{\delta}}{\beta_\theta}\right] \tag{72}$$

$$\zeta_{\theta^*} = \frac{1}{\mathcal{M}_{o,e}}\iota_{\theta^*} b_{\theta^*}\left(1 - \mathbb{E}_{pc}\left[\frac{\mu_\theta - \mathcal{M}_\theta}{\mu_\theta}\right]\frac{\mu_{\theta^*}}{\mu_{\theta^*} - \mathcal{M}_{\theta^*}}\right)\frac{\bar{\epsilon} - \bar{\delta}}{\bar{\delta}}\mathbb{E}_{pc}\left[\frac{1}{\mu_\theta}\right]$$

□

B.8 Sketch of Proof of Proposition 6

First, I formulate the planner's problem. Under the assumptions stated in the text, the social planner's allocation maximizes:

$$\begin{aligned}
\mathcal{L} = \quad & \max_{M,\theta^*,\{C_s,,c_{s,\theta}n_{s,\theta},n_{s,o},n_{s,e}\}_s} \quad \sum_s \tfrac{L_s}{L} U(C_s, \bar{N}_s) \\
& + \sum_s \zeta_s \left[C_s - M\left(\int_{\theta^*}^{\infty} c_{s,\theta}^{(\sigma-1)/\sigma} dG(\theta)\right)^{\sigma/(\sigma-1)}\right] \\
& + \sum_s \lambda_s \left[1 - M\left\{\Psi_{s,e}(\tfrac{n_{es,}}{L_s\bar{N}_s}) + \int_{\theta^*}^{\infty}[\Psi_{s,\theta}(\tfrac{n_{s,\theta}}{\bar{N}_s}) + \Psi_{s,o}(\tfrac{n_{s,o}}{L_s\bar{N}_s})]dG(\theta)\right\}\right] \\
& + \int_{\theta}^{\infty} \lambda_s \left[c_{s,\theta} - A_\theta \textstyle\prod_s (n_{s,\theta})^{\alpha_s}\right] \\
& + \lambda_e[f_e > \textstyle\prod_s (\tfrac{n_{s,e}}{L_s})^{\alpha_s}] + \lambda_o[f_o > \textstyle\prod_s (\tfrac{n_{s,o}}{L_s})^{\alpha_s}]
\end{aligned}$$

The proof follows exactly the same steps as the proof for theorem 1. First, focusing on the first order conditions for individual varieties, we again know that conditional on GE outcomes (utility, consumption, price and wage aggregates, entry, exit), the first order conditions only coincide when markdowns are constant across all firms within a labor market s. Then use the first order conditions with respect to M to show that the planner sets the (homogeneous) infra-marginal surplus of consumption across all varities equal to a productivity weighted average of wage-bill weighted infra-marginal employment surpluses across worker groups. This shows that labor market power has to be equal in all markets for the first order condition to coincide with the free entry condition in the decentralized economy, conditional on the cutoff, utility and market price aggregates. The same argument as in theorem 1 can then be used to show that the planner's first-order condition with respect to the cutoff θ^* exactly corresponds to the zero profit condition, conditional on utility and market price aggregates. Finally, the first order conditions with respect to C_s and N_s in combination with the variety specific first order conditions shows

that quantities are indeed the same across varieties, only when labor market power is homogeneous across both firms and labor markets.

C Extensions

C.1 Local Labor Markets

I show how to extend the model to account for local labor markets. Final good firms sell output in a national output market, and hire labor in "local labor markets". Labor markets are local in the sense that workers are immobile across regions $r \in \mathcal{R}$.

C.1.1 Environment and Equilibrium

Preferences, product demand, and labor supply Each region r is populated by a mass L_r of households and the total mass of households in the economy is given by $L = \sum_r L_r$. Per-capita utility from consumption C and labor supply N depends on consumption of available varieties (common across all regions) and labor supply to individual jobs (specific to regions):

$$1 = \sum_{r' \in \mathcal{R}} \int_{\theta \in \Theta^{r'}} \Upsilon_{(\theta,r')}\left(\frac{c_r(\theta,r')}{C^r}\right) dM^C(\theta,r'), \tag{73}$$

$$1 = \int_{\omega \in \Omega^r} \Psi_{(\omega,r)}\left(\frac{n(\omega,r)}{N^r}\right) dM^E(\omega,r), \tag{74}$$

In each region r, workers maximize utility:

$$\max_{C^r, N^r, c_r(\theta,r'), n(\omega,r)} C^r + \frac{(N^r)^{1+1/\varphi}}{1+1/\varphi},$$

subject to (73), (74), and the regional budget constraint

$$\sum_{r' \in \mathcal{R}} \int_{\theta \in \Theta^{r'}} p(\theta,r') c_r(\theta,r') dM^{r',C}(\theta) = \int_{\omega \in \Omega^r} w(\omega,r) n(\omega,r) dM^{r,E}(\omega).$$

The budget constraint anticipates that consumption varieties sell at the same price across all markets, and that due to free entry, there will be no profits distributed in equilibrium.

The homotheticity of the demand system implies that all regions consume the same relative per-capita quantities of available consumption varieties, so per-capita demand faced by final good producer (θ, r') is characterized in terms of an economy-wide output market competition index $\mathcal{P}$, and the relative demand it faces in its own local market.

$$p(r',\theta) = \mathcal{P} \Upsilon'_{(\theta,r')}\left(\frac{c_{r'}(\theta,r')}{C^{r'}}\right) Y,$$

where $Y = \sum_r Y_r$ is total per-capita GDP in across all regions and the economy-wide output competition index $\mathcal{P}$ solves:

$$1 = \sum_{r' \in \mathcal{R}} \int_{\theta \in \Theta^{r'}} \Upsilon_{(\theta,r')} \left((\Upsilon'_{(\theta,r)})^{-1} (\frac{p(\theta,r')}{\mathcal{P}}) \right) dM^C(\theta,r').$$

Labor supply to an employer ω in region r, in turn, is given by:

$$\frac{w(\omega,r)}{\mathcal{W}^r} = \Psi'_{(\omega,r)}(\frac{n(\omega,r)}{N^r}),$$

where the wage index $\mathcal{W}^r$ is region specific and solves

$$\frac{1}{\mathcal{W}^r} = \frac{1}{N^r} \int \Psi'_{r,\omega}(\frac{n_r(\omega)}{N^r}) \frac{n_r(\omega)}{N^r} dM^{r,E}(\omega).$$

Output markets In each region r, the economy resembles that described in the main text: Prospective entrants in the final goods sector purchase local entry goods at price $p_{e,r} f_{e,r}$ from local entry good producers, receiving a type realization θ from a cdf $G_r(\theta)$. To set up production, a producer has to incur overhead costs given by $p_{o,r} f_{e,r}$ in order to start production. Hiring $n_r(\theta)$ units of labor, a final good producer produces $y_r(\theta) = A_\theta n_r(\theta)$ units of final goods.

Final goods sell in all markets, and the characterization of demand above implies that producers in each region choose a price, wages, and overall employment level to maximize profits given by:

$$\pi(\theta,r) = \max_{w(\theta,r), p_{r'}(\theta,r)} \sum_{r'} L^{r'} p_{r'}(\theta,r) c_{r'}(\theta,r) - L^r w(\theta,r) n(\theta,r),$$

subject to per-capita product demand, labor supply, and technology:

$$p(\theta,r) = \mathcal{P} \Upsilon'_{(\theta,r)}(\frac{c_{r'}(\theta,r)}{C^{r'}}) Y^{r'}$$

$$\frac{w(r,\omega)}{\mathcal{W}^r} = \Psi'_{(\theta,r)}(\frac{n(\theta,r)}{N^r}) Y_r,$$

$$\sum_{r'} L^{r'} c_{r'}(\theta,r) = L^r A_\theta n(\theta,r).$$

In equilibrium, costless trade in final goods across regions implies that $\frac{c_{r'}(\theta,r)}{C^{r'}} = \frac{c(\theta,r)}{C}$ for all r', so firms offer the same price in each market satisfying:

$$p(\theta,r) = \frac{\mu(\theta,r)}{\mathcal{M}(\theta,r)} \frac{w(\theta,r)}{A_\theta},$$

where the markdown is given by $\mathcal{M}(\theta, r) = \frac{\beta(\theta,r)}{\beta(\theta,r)+1}$, with demand elasticity $\beta(\theta, r) \equiv \frac{\Psi'_{(\theta,r)}(\frac{\theta,r)}{N^r})}{\Psi''_{(\theta,r)}(\frac{n(\theta,,r)}{N^r})\frac{n(\theta,r)}{N^r}}$, and the markup is given by $\mu(\theta, r) = \frac{\sigma(\theta,r)}{\sigma(\theta,r)-1}$, with demand elasticity given by $\sigma(\theta, r) \equiv -\frac{\Upsilon'_{(\theta,r)}(\frac{c(\theta,r)}{C^r})}{\Upsilon''_{(\theta,r)}(\frac{c(\theta,r)}{C^r})\frac{c(\theta,r)}{C^r}}$. Imposing that $Lc(\theta, r) = L^r A_\theta n(\theta, r)$, these equations pin down wages, prices, markdowns, markups, and employment allocations across all active firms in all regions.

As before, firms in each region produce if, and only if, variable profits exceed the cost of overhead:

$$(1 - \frac{\mathcal{M}(\theta, r)}{\mu(\theta, r)}) \sum_{r'} L^{r'} p(\theta, r) c_{r'}(\theta, r) \geq p_{o,r} f_{o,r}.$$

Under the assumption that variable profits are strictly increasing and continious in type θ in every region r, in each region r exist a cutoff θ^*_r so that entrants with type realizations $\theta \geq \theta^*_r$ produce, while those with realizations $\theta < \theta^*_r$ exit. Given a mass of entrants M^r, the mass of available consumption varities and jobs (θ, r) is given by $dM^C(\theta, r) = dM^E(\theta, r) = M_r g_r(\theta) 1_{\{\theta \geq \theta^*_r)} d\theta$.

Free entry in each region implies that expected profits upon entry equal the cost of the entry good:

$$\int_{\theta^*_r}^{\infty} \left[(1 - \frac{\mathcal{M}(\theta, r)}{\mu(\theta, r)}) \sum_{r'} L^{r'} p(\theta, r) c_{r'}(\theta, r) - p_{o,r} f_{o,r} \right] dG_r(\theta) = p_{e,r} f_{e,r}.$$

The price of entry and ovearhead goods is region-specific, and depends on local labor market conditions:

$$p_{e,r} = \mathcal{W}^r \Psi'_{(e,r)}(\frac{f_{e,r}}{N^r L^r}) Y_r,$$

$$p_{o,r} = \mathcal{W}^r \Psi'_{(o,r)}(\frac{f_{o,r}}{N^r L^r}) Y_r.$$

The mass of jobs in the entry and overhead sector in region r are given by: $dM^E(e, r) = M_r$ and $dM^o(o, r) = M_r(1 - G_r(\theta^*_r))$.

Equilibrium Given $\{L_r, f_{e,r}, f_{o,r}, G_r(\theta)\}_{r \in \mathcal{R}}$ and the specification of preferences, a decentralized equilibrium is defined by allocations $\{c(\theta, r), n(\theta, r), n(e, r), n(o, r)\}$, prices/wages $\{p(\theta, r), w(\theta, r), p(o, r), p(e, r)\}$, cutoffs $\{\theta^*_r\}_{r \in \mathcal{R}}$, and mass of entrants $\{M_r\}_{r \in \mathcal{R}}$ that solve the utility-maximization problem of households, the profit-maximization problems of producers, and that clear goods and labor markets.

C.1.2 Efficiency

Consider a benevolent social planner that directly chooses quantities $\{c(\theta,r), n(\theta,r), n(e,r), n(o,r)\}$, selection cutoffs θ^r and entrants $\{M_r\}_{r\in\mathcal{R}}$ so as to maximize:

$$\max_{C_r N_r, \{c(\theta,r),n(\theta,r),n(e,r),n(o,r)\}, \{M_r,\theta_r^*\}} \sum_{r\in\mathcal{R}} \lambda_r^{\mathcal{W}} \left[C^r - \frac{(N^r)^{1+1/\varphi}}{1+1/\varphi} \right]$$

subject to

$$\begin{aligned}
&\forall r, && 1 = \textstyle\sum_{r'\in\mathcal{R}} M_{r'} \int_{\theta_{r'}^*}^{\infty} \Upsilon_{(\theta,r')}\left(\frac{c_r(\theta,r')}{C^r}\right) dG_{r'}(\theta),\\
&\forall r && 1 = M_r \left[\Psi_{(e,r)}\left(\frac{n(e,r)}{N^r}\right) + \int_{\theta_r^*}^{\infty} \left[\Psi_{(\theta,r)}\left(\frac{n(\theta,r)}{N^r}\right) + \Psi_{(o,r)}\left(\frac{n(o,r)}{N^r}\right)\right] dG_\theta(r) \right],\\
&\forall (\theta,r) && \textstyle\sum_{r'} L_{r'} c_{r'}(\theta,r) = L_r A_\theta n(\theta,r),\\
&\forall r && L_r n_{o,r} \geq f_{o,r}, L_r n_{e,r} \geq f_{e,r},
\end{aligned}$$

where $\lambda_r^{\mathcal{W}}$ is the weight the planner assigns to region r. The following theorem provides a generalizes Theorem 1 to this economy.

Theorem 3. *Suppose $\varphi \to \infty$, $\forall r$, $N_r = \bar{N}$, $\forall(\omega,r)$ $\Psi_{(\omega,r)}(x) = a_{(\omega,r)} x^{(\beta+1)/\beta}$ and $\forall$ (θ,r), $\Upsilon_{(\theta,r)}(x) = b_{(\theta,r)} x^{(\sigma-1)/\sigma}$. Then the market allocation coincides with the allocation chosen by a social planner with utilitarian welfare weights $\lambda_r^{\mathcal{W}} = \frac{L_r}{L}$.*

Further, thought-experiments similar to to the ones analyzed in Section 3 inform key statistics that characterize distortions in a given observed allocation in a regional labor market r. The following lemma summarizes these results.

Lemma 5. *The margins of inefficiency in an observed allocation, in each labor market r, are characterized as follows:*

1. *In region r, employer θ is too small compared to another employer θ', if, and only if, $\frac{\mu_\theta}{\mathcal{M}_\theta} > \frac{\mu_{\theta'}}{\mathcal{M}_{\theta'}}$*

2. *In region r, entry is insufficient if, and only if, $\frac{\bar{\epsilon}_r}{\bar{\delta}_r} > \mathbb{E}^r_{pc}\left[\frac{\mathcal{M}_\theta}{\mu_\theta}\right]^{-1}$, where $\bar{\epsilon}_r = \mathbb{E}^r_{pc}[\epsilon_\theta]$ and $\bar{\delta}_r = \mathbb{E}^r_{wn}[\delta]$, and $\mathbb{E}^r_{pc}[x] \equiv \int_{\theta_r^*}^{\infty} \frac{p(\theta,r)c(\theta,r)}{\int_{\theta_r^*}^{\infty} p(\theta,r)c(\theta,r)dG_r(\theta)} x_\theta dG_r(\theta)$.*

3. *In region r, selection is too weak if, and only if, $\bar{\epsilon}_r - \epsilon_{\theta_r^*} + \frac{\mathcal{M}_{\theta_r^*}}{\mu_{\theta_r^*}}\left(\delta_{\theta_r^*} - \bar{\delta}_r\right) + \left(1 - \frac{\mathcal{M}_{\theta_r^*}}{\mu_{\theta_r^*}}\right)(\delta_{o,r} - \bar{\delta}_r) > 0$.*

C.2 Heterogeneous Worker Types

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

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

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:

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

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:

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

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:

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

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.[30] The following 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* $a_\omega, b_{\omega'} \in \mathbb{R}^+$.

Proof. See Appendix B.8. □

Proposition 6 shows tthat efficiency in an economy with heterogeneous worker types and Cobb-Douglas production technologies requires that firms have a homogeneous degree of labor market power in all labor markets. 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: $\sum_s \alpha_s \mathcal{M}_{s,\theta} \neq \prod_s \mathcal{M}_{s,\theta}^{\alpha_s}$.

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.

[30] For the detailed description of the planner's problem see See Appendix B.1.

C.3 Variable Elasticity (VES) labor supply

The Kimball labor supply system embedded in the benchmark model has two advantages: First, it is homothetic. Therefore, it has a natural microfoundation based on aggregating individual labor supply decisions of workers, and can be easily embedded into richer models with heterogeneous workers or industries. Second, it allows each firms' markdown and pass-through to vary as a function of firm size, 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)). While the quantitative welfare implications of monopsony under this alternative preference specification differ, I show that efficiency remains tied to isoelastic labor supply.

Assume that the labor disutility N experienced by a household supplying $\{n_{\omega'}\}_{\omega'}$ hours is 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 7. *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.