Price discrimination in Supply Chains*

Luca Lorenzini — UCLA Anderson

Antonio Martner — UCLA & Central Bank of Chile

[Preliminary version: March 20, 2025]

Abstract

We study how price discrimination in supply chains affects resource allocation, economic rent distribution, and welfare. Using Chilean firm-to-firm transaction data, we document prevalent quantity discounts (second-degree price discrimination) alongside buyer-specific pricing (third-degree). We develop a supply chain model in which firms engage in price discrimination as sellers and are also price-discriminated against as input buyers. Relative to linear price setups, price discrimination increases output per firm but reduces firm entry by distorting the distribution of firm profits. Quantitatively, we find that the output increase effect dominates, resulting in higher welfare when firms can price discriminate compared to linear price setups. Our calibrated model, based on Chilean transaction data, suggests that the negative welfare effects of markups under price discrimination are approximately 15% smaller relative to linear pricing, indicating standard linear pricing setups may overestimate the welfare costs of market power.

*We thank Hugo Hopenhayn for excellent guidance, we are deeply indebted to him. Michael Rubens, John Asker, David Baqaee, and Federico Huneeus provided highly valuable comments. The views expressed are those of the authors and do not necessarily represent the views of the Central Bank of Chile or its board members. Authors emails: lucalorenzini@ucla.edu , amartner@ucla.edu

1 Introduction

Price discrimination in firm-to-firm transactions, while a well-documented phenomenon that shapes firms' contracts, remains often overlooked when analyzing the aggregate effects of market power. Competition between firms has been argued to be essential for efficient market outcomes, with competitive pressures driving prices toward marginal costs and benefiting buyers. Standard models of market power typically assume linear pricing, where all buyers face identical per-unit prices, and conclude that market power reduces welfare by decreasing output, distorting the allocation of production factors, and altering economic rent distribution, which also distorts firm entry.

However, linear pricing frameworks fail to capture observed pricing strategies in supply chains, where firms simultaneously act as price-discriminating sellers and price-discriminated buyers. Under price discrimination in supply chains, the average price is not fully allocative, and therefore, average markups are no longer a sufficient statistic to measure inefficiencies. If firms can price discriminate, market power will still harm welfare, but it might generate different allocations and surplus transfers that could help dampen market power's welfare costs, as suggested by Cowan (2016).

Using detailed firm-to-firm transaction data from Chile, we find that price dispersion is pervasive, where we cannot reject the presence of price discrimination in approximately 70% of transactions. How do firms set prices in firm-to-firm transactions? What are the welfare costs of price discrimination in supply chains compared to linear pricing scenarios? And what policies could effectively mitigate potential distortions? This paper addresses these questions by investigating the role of price discrimination in firm-to-firm transactions and quantifying its effects on resource allocation, rent sharing, and aggregate welfare.

Using our data, we find that quantity discounts, consistent with second-degree price discrimination, account for the largest share of observed price variations relative to third-degree price discrimination. While we do not provide a structural decomposition of both price discrimination degrees, when controlling for price-quantity menus, we find evidence suggesting that sellers offer buyer groups specific price-quantity menus, as the price variation is minimized when conditioning on seller-product-buyer-quantity combinations.

To analyze the aggregate implications of firm-to-firm price discrimination, we develop a general equilibrium model with heterogeneous firms in supply chains and endogenous entry, where sellers engage in price discrimination. We assume that sellers know the dis-

tribution of buyer types (productivities) but cannot identify specific firm types. Sellers are aware of a subset of buyer firm observables and endogenously choose to price discriminate based on buyers' productivity distributions (second-degree) and on firm observables they know (third-degree).

As sellers, firms offer optimal nonlinear price-quantity menu contracts conditional on firm observables. As buyers, firms are subject to the same price discrimination practices by their sellers. This recursive structure captures how price discrimination operates in supply chains, where firms must optimize their pricing strategies while accounting for how they themselves are being price discriminated against.

Under Pareto-distributed firm productivity and relying on mechanism design tools, we show that the optimal price-quantity menu takes the form of a two-part tariff consisting of a flat fee and a linear price component. The flat fee is set to extract buyers' surplus but does not have allocative implications; it will be innocuous to the contract quantities, while the linear price determines quantity allocations, which we refer to as the allocative markup.

In general equilibrium, two distortions are introduced with price discrimination relative to the planner's outcome. First, the allocative markup acts as a tax on output, reducing firms' input demands and distorting both buyers' and sellers' output. Second, a firm profit distortion occurs, where low-productivity firms have their surplus disproportionately extracted by sellers, decreasing firms' profits and effectively functioning as a tax on entry.

In linear markup setups, where firms exert market power but cannot price discriminate, the output tax is present, but all firms' profits have a linear surplus extraction from buyers. Relative to linear pricing setups, our price discrimination setup mitigates the allocative distortion, where each firm has, in expectation, higher output. Relative to a linear pricing equilibrium, high-productivity firms earn higher profits because they retain information rents, which restricts sellers' ability to capture a larger share of their surplus. In contrast, low-productivity buyers face greater profit distortions, as sellers extract more surplus from them than under linear pricing. This asymmetry in surplus extraction alters the profit distribution across firms, decreasing firm entry relative to linear price setups. As a result, price discrimination generates a trade-off between firm-level output and firm entry, with aggregate output and welfare implications.

We further characterize optimal policy interventions that restore efficiency. The planner can replicate the efficient benchmark in a decentralized price discrimination econ-

omy through a combination of differentiated output subsidies and redistributive transfers across firms. A firm-specific output subsidy corrects allocative distortions, while a lump-sum transfer or tax mitigates surplus extraction inefficiencies. Relative to linear pricing, price discrimination increases per-firm output but reduces firm entry, leading to nontrivial welfare effects.

We calibrate the model using moments from the firm size distribution we observe in the data and parameter values sourced from the literature. We do not target the unit price-quantity discounts observed in the data, allowing us to use them as an out-of-sample test of the model's ability to rationalize this pricing pattern. We find that the model replicates both the shape of the unit price function, which is well characterized by a two-part tariff, and the magnitude of quantity discounts observed in firm-to-firm transactions.

Our quantitative results indicate that the price discrimination equilibrium improves efficiency relative to the linear pricing model, but output is still below the efficient allocation. Relative to linear pricing, price discrimination reduces firm entry by approximately 10–15%, depending on input substitutability, but increases average output per firm. As a result, aggregate output is 15% higher under price discriminationwhen input substitutability is low and approximately 0.6% higher when input substitutability is high. However, despite these efficiency improvements, the aggregate output of the economy remains 25–50% below its efficient allocation level. The extent of these distortions depends on the firm productivity distribution shape parameter and input substitutability. In particular, when firms exhibit a high elasticity of substitution, both the price discrimination and the linear economy converge toward efficiency, as input distortions become less relevant.

We then evaluate the role of policy interventions in mitigating the inefficiencies associated with price discrimination. We analyze three counterfactual policy experiments: (i) a lump-sum transfer from upstream to downstream firms to offset profit extraction, (ii) an output subsidy that restores marginal cost pricing conditional on firm entry, and (iii) a uniform output subsidy applied to both upstream and downstream firms. Our findings suggest that an undifferentiated uniform output subsidy applied to all firms effectively restores efficiency, closing the 25-50% gap in output relative to the efficient level. This policy functions as if the economy were operating under linear pricing and corrects both input and profit distortions while aligning firm incentives with the efficient benchmark.

By contrast, an entry subsidy that offsets profit extraction without a production subsidy reduces welfare relative to the price discrimination economy without policy inter-

vention. Similarly, an output subsidy that restores marginal cost pricing conditional on entry improves efficiency, but by less than the undifferentiated uniform output subsidy.

These results highlight the fundamental trade-offs between price discrimination and linear pricing structures in supply chains. While price discrimination improves firm expected output relative to linear pricing, it introduces profit distortions that affect firm entry decisions. Our results suggest that standard linear pricing setups may overestimate the welfare costs of market power.

Related Literature

This paper connects to the literature on firm heterogeneity, market power, and optimal pricing strategies by offering a quantitative framework to explore the macroeconomic effects of price discrimination in supply chains. We engage with three strands of literature.

First, we build upon research on price discrimination in intermediate goods markets by documenting its prevalence across firms and quantifying its welfare implications. Our work complements Burstein et al. (2024), who study input price dispersion across buyers and its misallocation effects, by endogenizing firms' pricing behavior and considering both third-degree and second-degree price discrimination. We further examine the welfare consequences of these pricing strategies. Our empirical contribution is to provide new evidence of significant quantity discounts in firm-to-firm transactions. The theoretical contribution is to develop a novel general equilibrium model in which firms engage in a mechanism design problem with one another.

Second, we engage with the literature on production networks by incorporating more realistic pricing mechanisms that better align with empirical patterns. We build on the aggregate market power frameworks of Edmond et al. (2023) and integrate entry dynamics from Baqaee and Farhi (2020), who analyze markup distortions in production networks. By introducing price discrimination, we modify how these distortions propagate through supply chains.

Third, we add to studies on market power by demonstrating how price discrimination can partially mitigate welfare losses from markups through improved resource allocation. While De Loecker et al. (2021) and Boehm et al. (2024) examine the welfare implications of market power in models with firm dynamics, and Hsieh and Klenow (2014) analyze the misallocation effects of resource distortions, we show that incorporating observed pricing practices reduces estimated welfare losses from market power. In this respect, our findings align with Bornstein and Peter (2024), highlighting that price discrimination

is a critical factor when assessing the aggregate implications of firm-level market power.

2 Descriptive Evidence

Using Chilean Internal Revenue Service firm-to-firm transactions data, we examine price variations at the seller-product level. We find that nonlinear pricing is prevalent, where quantity discounts, consistent with second-degree price discrimination, account for the largest share of observed prices.

2.1 Data Description

The Chilean Internal Revenue Service (IRS) data contains records of firm-to-firm transactions, including product descriptions, prices, and quantities, which can be merged with firms' accounting variables, including total revenue, labor costs, materials costs, and capital expenditure.

We use data from the universe of Chilean firm-to-firm value-added tax invoices collected by the Chilean IRS.[1] For each transaction-specific invoice, we observe seller and buyer IDs, a description of every product within the invoice, and the corresponding price and quantity. We analyze data from 2018 to describe the dataset and its main characteristics. That year, 537,521 seller IDs and 3,398,323 buyer IDs traded a total of 60,029,741 different products for a total of 1.24 billion products traded.

Our focus for this section is on the raw data, while some aggregation and industry conditioning have proven helpful in uncovering certain characteristics, as Burstein et al. (2024) showed with the same database, we center our analysis on the most granular data available without excluding any industry. We define our unit of observation as each line item within invoices between two tax identifiers[2]. In each line, we observe the transacted product "detail" which describes the product (for example, blue paint brand XX 3 gallons)

[1]This study was developed within the scope of the research agenda conducted by the Central Bank of Chile (CBC) in economic and financial affairs of its competence. The CBC has access to anonymized information from various public and private entities, by virtue of collaboration agreements signed with these institutions. To secure the privacy of workers and firms, the CBC mandates that the development, extraction and publication of the results should not allow the identification, directly or indirectly, of natural or legal persons. Officials of the Central Bank of Chile processed the disaggregated data. All the analysis was implemented by the authors and did not involve nor compromise the Servicios de Impuestos Internos de Chile.

[2]Not necessarily firms as some tax IDs do not report hiring workers, purchasing intermediate inputs, or capital expenditure.

and is often seller firm-specific; we refer to this variable as product. We observe each product's unit price and quantity. Importantly, in the majority of transactions, shipping costs are recorded as a separate product on the invoice, implying that the unit prices in our analysis do not include shipping costs. One of the main challenges of the data is that we lack standardized information on product units, a limitation we address by treating products as firm-specific.

We find that firms in Chile have a clear partition on firms' buyers: 79% of firms weighted by sales sell all their output either to only other firms (67%) or to only final consumers (12%). As we can combine firm-to-firm transaction data with firms' accounting variables, we build an indicator variable that takes the value of 0 if all firm sales go to final consumers and 1 if sales go only to other firms, and we weigh the indicator by firm sales.

Table 1: Firms sales partition

Sector (GDP sales share)	All to final consumer	All to other firms
Firm population (100%)	0.12	0.67
Agriculture (2%)	0.05	0.60
Mining (1%)	0.27	0.08
Manufacturing (15%)	0.06	0.69
Utilities (3%)	0.20	0.52
Construction (8%)	0.02	0.89
Retail and Wholesale (32%)	0.09	0.69
Transport and ICTs (10%)	0.16	0.68
Financial Services (18%)	0.18	0.68
Real Estate Services (1%)	0.25	0.38
Business Services (7%)	0.09	0.81
Personal Services (2%)	0.69	0.10

Notes: Exports are excluded. The remaining 16% of sales shares for the firm population are firms that sell to both final consumers and other firms. We observe firm-to-firm sales an fims total sales, we compute the sales to consumer as the residual between both. For 2% of firms, we get negative sales to consumers and exclude them from this graph.

As shown in Table 1, there is heterogeneity across sectors, though the partition between firms selling to final consumers and other firms is present across all sectors.

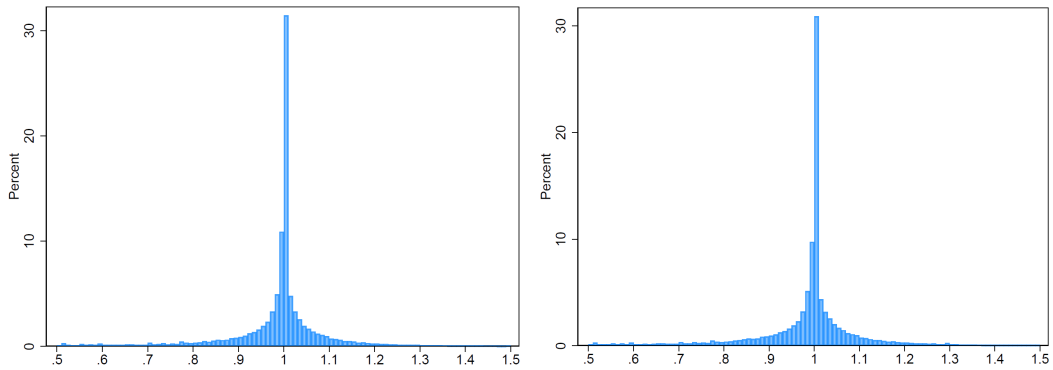
2.2 Price Determinants

We observe substantial price variations for a given seller i and product g (the "detail" variable in the invoice) within a month. Following Burstein et al. (2024), we construct a price dispersion measure α_{ijg} for June 2018, the month with more transactions in 2018, a year where Chile did not experience any major shock. We divide unit prices observed for each product g transaction from seller i to buyer j by the mean price across seller i and product g. We repeat the same exercise for June 19th, 2018, the day with more transactions

of that month, to be sure our results are not driven by monthly-specific demand and supply shocks.

Figure 1: Price dispersion

Panel A. June 2018 Panel B. June 19th 2018



Notes: The figure reports the distribution of the log of demeaned price for the month of June 2018. We exclude seller-product pairs with only one transaction.

The variance of $\ln(\alpha_{ijg})$ is 0.65 monthly and 0.61 daily, and 29% of transactions in both cases show no price dispersion, as illustrated in Figure 1. For 71% of transactions, we cannot reject that firms engage in some form of price discrimination. While first-degree price discrimination, setting prices along the complete demand curve, is unlikely, we investigate whether firms engage in second-degree price discrimination through nonlinear pricing (quantity discounts) or third-degree price discrimination (charging different prices to different buyers). To analyze price variance determinants, we estimate the following regression:

To analyze price variance determinants, we estimate the following regression:

$$\ln\ p_{ijgm} = \ln G + \ln\ \psi + \ln\ \xi_m + \ln\ \omega_{ijgm}$$

where G is the mean across all sellers, buyers, products and months, ψ represents different fixed effects interaction strategies, ξ_m is a month fixed effect, and ω_{ijgm} is the residual, interpreted as match-specific unobservables for each seller-buyer-product-month (i, j, g, m) combination. The interaction between seller and product fixed effects is needed to control for product unit heterogeneity.

Table 2 shows that both second and third-degree price discrimination explain price variance beyond seller-product fixed effects. Based on the residual variance and partial

R^2, quantity variation accounts for a larger share of unexplained variation than buyer heterogeneity. This suggests that second-degree price discrimination (quantity-based pricing) dominates third-degree price discrimination (buyer-based pricing).

Table 2: Price observables determinants

	(1)	(2)	(3)
FE Seller-product	Yes	No	No
FE Seller-product-Buyer	No	Yes	No
FE Seller-product-quantity	No	No	Yes
FE month	Yes	Yes	Yes
Residual variance	0.46	0.45	0.38
Observations	1.2 billions	1.2 billions	1.2 billions
Adjusted R^2	0.8835	0.8883	0.9035
Partial adjusted R^2	–	0.041	0.172

Notes: Partial adjusted R^2 is computed by the ratio of the R^2 difference with respect to regression(1) to $1-R^2$ using regression (1) R^2

We perform a price variance decomposition using observable variables to separately quantify the variance shares attributed to second and third-degree price discrimination. We estimate the following equation:

$$\ln\ p_{ijgm} = \ln G + \ln\ \psi_{ig} + \ln\ \theta_j + \ln\ \gamma_q + \ln\ \epsilon_{ijgm}$$

where ψ_{ig} represents seller-product interacted fixed effects, θ_j represents buyer fixed effects, and γ_q represents quantity fixed effects. We construct quantity fixed effects by dividing observed quantities into 1000 quantile bins. Table 3 shows that the quantity fixed effect associated with second-degree price discrimination explains a larger share of price variation than the buyer fixed effect associated with third-degree price discrimination.

Table 3: Price variance decomposition

	$\frac{\text{var}(\theta_j)}{\text{var}(\theta_j+\gamma_q)}$	$\frac{\text{var}(\gamma_q)}{\text{var}(\theta_j+\gamma_q)}$	$\frac{2\text{cov}(\theta_j,\gamma_q)}{\text{var}(\theta_j+\gamma_q)}$
Variance decomposition	0.39	0.77	-0.16
Lower bound	0.23	0.61	
Upper bound	0.55	0.93	

Notes: The lower and upper bounds are computed by adding and subtracting the covariance share term to the variances shares.

Given that transaction quantities explain the larger share of price variance in our data, we focus on evaluating and documenting the presence of quantity discounts (second-degree price discrimination).

2.3 Evidence of Nonlinear Pricing

Treating observed prices as equilibrium outcomes of seller-buyer contracts, we find a negative correlation between unit prices and transaction quantities. This analysis does not estimate demand-price elasticities, which would be incorrect due to endogeneity issues, but rather documents the existence of nonlinear pricing using realized transaction data.

We begin by regressing log prices on log quantities with interacted seller-product-buyer fixed effects. We focus on bilateral relationships between firm pairs for specific products, which approximates contract agreements. Price variation within these seller-product-buyer combinations indicates quantity-based pricing. We estimate the following regression:

$$\ln\ p_{ijgm} = \beta_0 + \beta_1 \ln\ q_{ijg} + \ln\ \psi_{ijg} + \ln\ \epsilon_{ijgm}$$

where q_{ijg} represents the observed quantity in each transaction. Table 4 shows that, on average, doubling the quantity purchased reduces the unit price by 25%. In column (2), we re-estimate this relationship using only products in the top five percent of the price distribution. Fixed costs associated with packaging, bureaucratic expenses, and personalization[3]—which could rationalize quantity discounts through increasing returns—should be less relevant for these high-price products. The results confirm that quantity discounts

[3]Recall that price data do not include transportation costs.

persist even at the upper end of the price distribution, suggesting that they remain significant for products where fixed costs play a smaller role. This effect is consistent across sectors, including manufacturing firms (25% of transactions) and retail and wholesale firms (61% of transactions). In Appendix A, we estimate the regression separately for 11 aggregated sectors, finding that the coefficient ranges from -0.161 in Manufacturing to -0.773 in Real Estate Services.

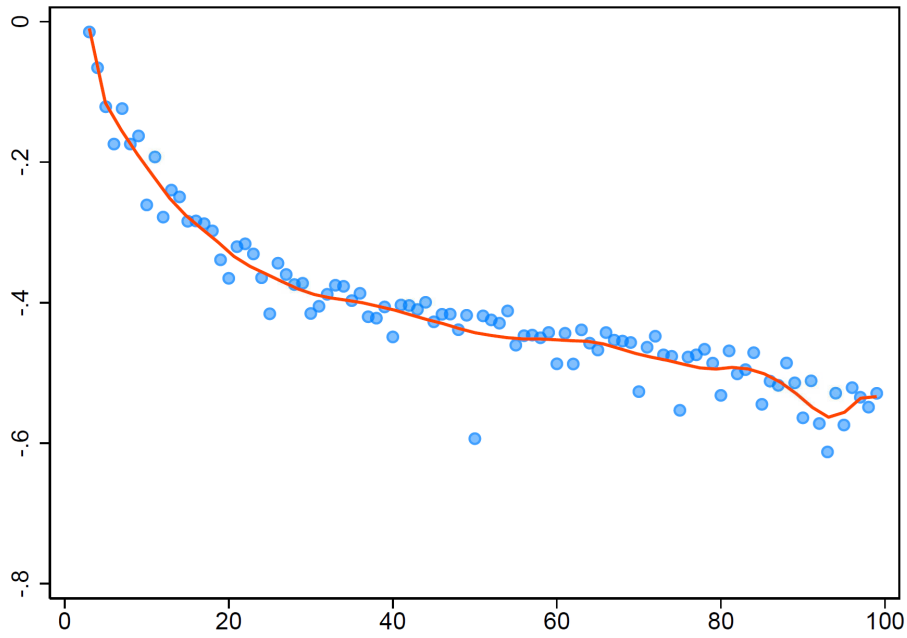
Table 4: Quantity discounts regression results

	(1)	(2)	(3)	(4)
ln q	-0.248	-0.188	-0.213	-0.224
Standard errors	(0.00002)	(0.00025)	(0.00005)	(0.00004)
FE Seller x product x buyer	Yes	Yes	Yes	Yes
FE month	Yes	Yes	Yes	Yes
High price products	No	Yes	No	No
Ex-manufacturing products	No	No	Yes	No
Ex-Retail and wholesale	No	No	No	Yes
Observations (millions)	1246	61	932	480
Adjusted R2	0.8981	0.7007	0.7561	0.7854

Notes: High product prices are defined as products with a price above the 95 percentile of the price distribution observed for year 2018.

To analyze the quantity discount scheme with higher precision, we estimate a regression using unit-step quantity bins to evaluate how price discounts vary across the quantity distribution. We construct 100 bins with unit-width intervals from 1 to 100 (for example, bin 2 includes all quantities greater than 1 and less than or equal to 2), covering 98% of transactions.

Figure 2: Quantity discount by quantity bins



Notes: The line is a order 10 polynomial fit of the fixed effects and the one unit step bins.

Figure 2 plots the estimated fixed effects and includes a polynomial fit line. Quantity discounts are largest for initial units, and while the price-quantity relationship remains negative throughout, the marginal discount diminishes for larger quantities. To test whether the data aligns with firm-specific quantity-price menus, we examine price variation within seller-product pairs at fixed quantities, computing price standard deviations and medians for pairs with multiple transactions. If pricing strategies were characterized by price-quantity menus, residual price variation should drop to zero when conditioning on the contract. We construct a coefficient of variation by dividing the standard deviation by median prices and calculate its distribution moments. Table 5 presents these results across different fixed effects specifications. The coefficient of variation decreases as we narrow the analysis to within seller-buyer-product-quantity combinations. While the coefficient of variation remains positive for the top of the coefficient distribution, the evidence supports the presence of quantity-price menus.

Table 5: Coefficient of variation

	coefficient of variation				
	p10	p25	Median	p75	p90
Seller - product - month	0.00	0.04	0.13	0.54	4.11
Seller - product - month - quantity	0.00	0.01	0.08	0.35	2.06
Seller - product - month - quantity - Buyer	0.00	0.00	0.00	0.04	0.45

Notes:

Our findings can be summarized in four points. First, firms predominantly sell either to final consumers or to other firms. Second, firms charge different prices for the same product, indicating price discrimination. Third, second-degree price discrimination (quantity discounts) explains a larger share of the observable price variance than buyer-based price dispersion. Fourth, doubling quantities reduces unit prices by 25% on average, with larger reductions occurring at lower quantities. The next section develops a theoretical framework that accounts for these four findings and moves towards endogenizing firms' pricing behavior.

3 A Model of Price Discrimination in Supply Chains

Based on the prevalence of nonlinear pricing observed in the data, we develop a model of supply chains in which firms engage in nonlinear pricing strategies with one another. We begin by describing the model environment and briefly discussing both the efficient benchmark and the equilibrium under linear pricing. Next, we present the model incorporating nonlinear pricing and provide counterfactual analyses comparing economic outcomes under nonlinear pricing with those under the efficient benchmark and linear pricing. The objective is to elucidate and quantify the inefficiencies arising from these pricing schemes. The economy is closed, and there is no aggregate uncertainty. Firms are infinitesimal so strategic interactions are ruled out. We focus on characterizing the steady state.

3.1 Environment

Industry Structure

There is a representative consumer who has preferences over a final consumption good, supplies labor, which is the only factor of production and owns firms. Motivated by Table 1, we model the supply chain as a two-layer structure, where downstream firms sell exclusively to the final good producer, while upstream firms engage only in transactions with other firms (either upstream or downstream). Both layers demand labor (details to follow). Figure 3 illustrates the assumed production network structure of the economy.

Within each layer, there exists a continuum of firm types, and within each type, a continuum of firms. For notation, we index downstream firm types by z and upstream firm types by ζ.

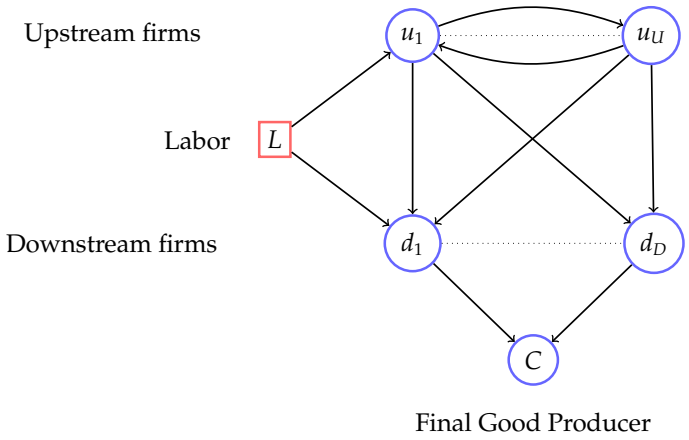
The final good is produced by a perfectly competitive firm using a constant elasticity of substitution (CES) aggregator over a continuum of goods produced by downstream firms. The CES aggregator is given by:

$$Y = \left(\int_{z_0} y(z)^{\frac{\sigma-1}{\sigma}} M_z \mu(z)\, dz \right)^{\frac{\sigma}{\sigma-1}},$$

where σ represents the elasticity of substitution across dowstream firm's varieties, $y(z)$ denotes the consumption of variety z, and $\mu(z)M_z$ represents the total mass of that variety. Note that Y corresponds to GDP in this economy.

To model industry dynamics, we extend the frameworks of Hopenhayn (1992) and Melitz (2003) to a two-layer structure. Specifically, there exists an unbounded pool of prospective entrants that are ex-ante identical. To enter, firms must pay a sunk cost c_e in units of labor. Upon entry, firms draw their type from a common distribution $G(\gamma)$, where $\gamma \in \{z, \zeta\}$. Firms then choose whether to enter the upstream or downstream sector. Let M_z denote the mass of firms in the downstream sector and M_ζ the mass of firms in the upstream sector, both of which are endogenous equilibrium objects.

Figure 3: Supply Chain Structure



Production

Both upstream and downstream firms produce using a CES production function that aggregates labor and a bundle of intermediate goods sourced from upstream firms:

$$q(\gamma) = \gamma \left(\alpha l(\gamma)^{\frac{\eta-1}{\eta}} + (1-\alpha) m(\gamma)^{\frac{\eta-1}{\eta}} \right)^{\frac{\eta}{\eta-1}}, \quad \gamma \in \{z, \zeta\},$$

where $l(\gamma)$ represents labor input, $m(\gamma)$ denotes the intermediate input bundle, and η is the elasticity of substitution between labor and intermediate inputs. The intermediate input bundle itself is a CES aggregator.[4] Denote by $m(\gamma, \zeta')$ the quantity purchased by firm $\gamma \in \{z, \zeta\}$ from upstream firm ζ'. Then:

$$m(\gamma) = \left(\int_{\zeta_0} m(\gamma, \zeta')^{\frac{\sigma-1}{\sigma}} M_\zeta \mu(\zeta') d\zeta' \right)^{\frac{\sigma}{\sigma-1}}, \quad \gamma \in \{z, \zeta\}.$$

3.2 Equilibrium Notions

We describe two different notions of equilibrium. The first equilibrium notion considers a setting where firms are constrained to offering only linear prices. In the second, which

[4] We implicitly assume that the network is fully connected. This can be interpreted in two ways: either firms purchase from all potential suppliers, or firms face a discrete choice problem in which they select a single supplier to maximize indirect utility, subject to a logit taste shock. For a similar micro-foundation, see Fajgelbaum et al. (2011), or for a micro-foundation in the context of workers choosing which firm to work for, see Berger et al. (2022).

constitutes the main theoretical innovation of this paper, we assume that upstream firms are free to charge *any* price per quantity to their buyers, thereby facing a mechanism design problem. This means that upstream seller firms can charge nonlinear prices to other firms meanwhile paying nonlinear prices to other upstream firms for their intermediate input bundle.

In both equilibrium notions, downstream firms and the final goods producer continue to offer linear prices. Moreover, in both cases, we impose feasibility constraints to ensure that supply equals demand. Additionally, in an equilibrium with unrestricted entry, the discounted present value of expected profits must equal the cost of entry at both layers.

We begin by describing the standard equilibrium conditions that apply to both benchmarks. We then proceed to illustrate the problems faced by upstream and downstream firms in the two different scenarios.

Final Good Producer

The final good producer minimizes expenditure subject to its production technology:

$$\min_{\{y(z)\}} \int_{z_0} p(z)y(z)\mu(z)M_z dz \quad \text{s.t.} \quad Y \geq 1,$$

which yields the demand functions:

$$y(z) = \left(\frac{p(z)}{P}\right)^{-\sigma} Y,$$

where

$$P = \left(\int_{z_0} p(z)^{1-\sigma}\mu(z)M_z dz\right)^{\frac{1}{1-\sigma}}.$$

We normalize the final good price to one.

Feasibility

Feasibility requires that, for every good produced—whether upstream or downstream—demand does not exceed supply. Thus, for downstream goods, feasibility requires:

$$y(z) \leq q(z) \quad \forall z.$$

Similarly, for upstream goods:

$$\int_{z_0} m(z,\zeta)\mu(z)M_z dz + \int_{\zeta_0} m(\zeta',\zeta)\mu(\zeta')M_\zeta d\zeta' \leq q(\zeta) \quad \forall \zeta.$$

Moreover, labor demand cannot exceed labor supply. Without loss of generality for our counterfactual analysis, we normalize labor supply to one. Thus, labor feasibility requires:

$$\int_{z_0} l(z)\mu(z)M_z dz + \int_{\zeta_0} l(\zeta)\mu(\zeta)M_\zeta d\zeta + c_e(N_z + N_\zeta) \leq 1,$$

where N_z and N_ζ denote the mass of entrants in the downstream and upstream layers, respectively. In a steady-state equilibrium, the mass of firms in both layers remains constant. Thus, the mass of entrants must exactly offset the mass of exiting firms, implying:

$$N_\gamma = (1-\delta)M_\gamma, \quad \gamma \in \{z,\zeta\}.$$

Furthermore, let L denote the total amount of labor used in production by upstream and downstream firms:

$$L = \int_{z_0} l(z)\mu(z)M_z dz + \int_{\zeta_0} l(\zeta)\mu(\zeta)M_\zeta d\zeta.$$

Substituting this condition into the labor feasibility constraint, we obtain:

$$L + c_e(1-\delta)(M_z + M_\zeta) \leq 1.$$

Zero Profit Condition

In an equilibrium with unrestricted entry, the discounted present value of expected profits must equal the cost of entry in both layers. Let $\pi(\gamma)$ denote the profits of upstream and downstream firms, and w be the wage. Then:

$$\begin{aligned} \frac{1}{1-\delta}\mathbb{E}_z[\pi(z)] &= c_e w, \quad \text{For downstream firms} \\ \frac{1}{1-\delta}\mathbb{E}_\zeta[\pi(\zeta)] &= c_e w, \quad \text{For upstream firms} \end{aligned}$$

3.2.1 Linear Pricing Equilibrium

The first notion of equilibrium considers a setting in which firms are constrained to offering a single linear price under monopolistic competition. Firms demand inputs to minimize expenditure:

$$\min_{\{m(\gamma,\zeta')\},l(\gamma)} \int_{\zeta_0} p(\zeta')m(\gamma,\zeta')M_\zeta\mu(\zeta')d\zeta' + wl(\gamma), \quad \gamma \in \{z,\zeta\}, \quad \text{s.t.} \quad q(\gamma) \geq 1,$$

which yields the following input demand functions:

$$m(\gamma,\zeta') = \left(\frac{p(\zeta')}{p_m}\right)^{-\sigma} m(\gamma), \quad \gamma \in \{z,\zeta\},$$
$$m(\gamma) = \left(\frac{\alpha}{1-\alpha}\frac{p_m}{w}\right)^{-\eta} l(\gamma), \quad \gamma \in \{z,\zeta\},$$

where p_m is given by:

$$p_m = \left(\int_{z_0} p(\zeta)^{1-\sigma}\mu(\zeta)M_\zeta d\zeta\right)^{\frac{1}{1-\sigma}}.$$

The firm's marginal cost is then given by:

$$MC(\gamma) = \frac{1}{\gamma}c(w,p_m) = \frac{1}{\gamma}\left(\alpha^\eta w^{1-\eta} + (1-\alpha)^\eta p_m^{1-\eta}\right)^{\frac{1}{1-\eta}}.$$

Moreover, both upstream and downstream firms choose their prices to maximize total profits:

$$\max_{p(\gamma)} \left(p(\gamma) - MC(\gamma)\right) D(\gamma), \quad \gamma = z,\zeta$$

Where $D(\gamma)$ is total demand of type γ firm. We assume that firms are infinitesimal and have no effects on others firm's pricing strategies, hence, each firm sets its price as a constant markup over marginal cost:

$$p(\gamma) = \frac{\sigma}{\sigma-1}\frac{1}{\gamma}c(W,p_m).$$

Definition 1. *Decentralized Linear Pricing Equilibrium*
A decentralized equilibrium is a collection of firm prices $\{p(\zeta),p(z)\}$, wage w, and quanti-

ties $\{y(z), q(z), l(z), m(z,\zeta'), l(\zeta), m(\zeta,\zeta'), M_z, M_\zeta\}$ such that, given technology $\{z,\zeta\}$:

i) The final good producer demands downstream goods to minimize costs.

ii) Both upstream and downstream producers demand intermediate inputs and labor to minimize costs.

iii) Each price equals marginal cost times the markup $\frac{\sigma}{\sigma-1}$.

iv) Entrants earn zero expected profit.

v) Feasibility constraints hold with equality.

We solve in detail the problem where firms engage in monopolistic competition charging linear prices in Appendix C.3

Lemma 1. ***Efficient Benchmark***
The Efficient Benchmark is attained by a Linear Pricing Decentalized Equilibrium with an output subsidy equal to $\tau = \frac{\sigma-1}{\sigma}$ given to all firms in the economy and paid via lump sum tax on the final consumer.

The lemma is a special case of the more general result in Theorem 1 of Baqaee and Farhi (2020), and the proof is provided in Appendix C.2. Intuitively, efficiency requires compensating entrants for the value they generate for society. The value of an additional firm to society is the sum of consumer and producer surplus. Efficiency dictates that entry occurs until the cost of entry equals the expected marginal value of entry.

Since the entry cost paid is equal to expected profits, efficiency requires that expected profits be equated to the expected marginal value of entry (i.e., consumer plus producer surplus). If firms were to price at marginal cost, their profits would correspond to producer surplus. However, due to the downward-sloping demand curve (arising from the love of variety), firms must be allowed to charge a markup sufficient to incentivize the optimal level of entry.

Nevertheless, this markup distorts input choices by effectively acting as a uniform tax on production. To mitigate this distortion, an output subsidy is required to restore marginal-cost pricing, conditional on entry.

3.3 Non-Linear Pricing Equilibrium

The second equilibrium notion we analyze is the nonlinear pricing equilibrium. In this setting, firms are no longer constrained to offering a single linear price; instead, they can select a continuum of price menus to charge their buyers. By allowing for nonlinear pricing, we implicitly assume the absence of a secondary market where products can be repackaged and resold at zero cost (i.e., no arbitrage opportunity). Instead, we assume that resale activity is costly enough that it does not arise in equilibrium.

We assume that firms can distinguish between upstream and downstream buyers, allowing them to offer different contracts for type-z and type-ζ firms. In particular, we assume that firms can price discriminate based on whether their buyers belong to the upstream or downstream sector. This constitutes a form of third-degree price discrimination.

This setting departs from the standard mechanism design framework because upstream firms are not only choosing their own nonlinear pricing schemes but are also subject to the mechanism design choices of other firms.

To be consistent with empirical evidence and to obtain a tractable solution, we introduce the following assumption:

Assumption 1. Firm productivity follows a Pareto distribution with tail parameter κ. For output to be finite, we require the tail condition $\kappa > \sigma - 1$[5] .

As we will show later, this assumption generates nonlinear pricing patterns consistent with the empirical findings of Figure 2.

Upstream firms Nonlinear Pricing Problem

We proceed by applying the method of guessing and verifying.

Guess 1: The total transfer follows a two-part tariff:

$$P(\gamma, \zeta) = p(\zeta)m(\gamma, \zeta) + F_\gamma(\zeta), \quad \text{for} \quad \gamma \in \{z, \zeta\}.$$

[5]For instance, in the linear benchmark economy, output is proportional to average productivity i.e. $Y \propto \tilde{z} = \left[\mathbb{E}_z z^{\sigma-1}\right]^{\frac{1}{\sigma-1}}$. $z^{\sigma-1}$ is itself distributed according to a Pareto distribution with tail $\frac{\kappa}{\sigma-1}$. Thus, to have a finite average, $\frac{\kappa}{\sigma-1} > 1$

Guess 2: The revenue functions take the following form:

$$R(z) = q(z)^{\theta} A_z, \quad R(\zeta) = q(\zeta)^{\theta} A_\zeta, \quad \text{for some} \quad \theta, A_\gamma, \quad \gamma \in \{z, \zeta\}.$$

To verify this guess, we analyze the problem of an infinitesimal firm indexed by i with productivity ζ. Under these assumptions, the i-th firm's marginal cost remains constant and is given by:

$$c(\zeta) = \frac{1}{\zeta} c(w(\zeta), p_m(\zeta)), \qquad c(w, p_m) = \left(\alpha^{\eta} w^{1-\eta} + (1-\alpha)^{\eta} p_m^{1-\eta}\right)^{\frac{1}{1-\eta}},$$

where p_m is the CES index of the linear price in the two-part tariff:

$$p_m = \left(\int p(\zeta)^{1-\sigma} M_\zeta \mu(\zeta)\, d\zeta \right)^{\frac{1}{1-\sigma}}.$$

Moreover, the pricing problem of the downstream firm, as well as the input minimization problem of each firm, remains identical to that in the linear price benchmark since the flat fee does not distort decisions (i.e., it does not affect the first-order conditions).

Applying the Revelation Principle, the upstream firm's problem is to choose a direct mechanism $\{m(\gamma, \zeta), P(\gamma, \zeta)\}$, where $P(\gamma, \zeta)$ denotes the total transfer from firm type γ to the seller ζ. This choice is subject to i) The Incentive Compatibility (IC) constraint, ensuring that buyers self-select into the mechanism designed for them. ii) The Individual Rationality (IR) constraint, ensuring that buyers are willing to participate. Denote by $\Pi(\gamma)$ the total profits of a buyer-firm of type $\gamma = z, \zeta$ and define the total buyer surplus from purchasing quantity x from seller ζ as:

$$\tilde{\Pi}(\gamma, x) = \frac{d\Pi(\gamma)}{d\left(\mu(\zeta) M_\zeta\right)}\Big|_{m(\gamma,\zeta)=x}, \quad \gamma \in \{z, \zeta\}.$$

The upstream seller's problem is:

$$\max_{\substack{\{P(z,\zeta), m(z,\zeta)\}, \\ \{P(\zeta',\zeta), m(\zeta',\zeta)\}}} \Pi(\zeta) = \underbrace{\mathbb{E}_z \left[P(z,\zeta) - C\left(m(z,\zeta)\right)\right]}_{\text{Profits from Downstream firms}} + \underbrace{\mathbb{E}_{\zeta'} \left[P(\zeta',\zeta) - C\left(m(\zeta',\zeta)\right)\right]}_{\text{Profits from other upstream firms}}$$

subject to:

$$\text{(IC)} \quad m(\gamma,\zeta) \in \underset{m(\gamma,\zeta)}{\operatorname{argmax}} \tilde{\Pi}(\gamma, m(\gamma,\zeta)) \quad \forall \quad \gamma \in \{z,\zeta'\}$$
$$\text{(IR)} \quad \tilde{\Pi}(\gamma, m(\gamma,\zeta)) \geq 0 \quad \forall \quad \gamma \in \{z,\zeta'\}.$$

Under the proposed guesses, and using buyers profit expressions, the upstream firm's problem simplifies to (see Appendix for further details):

$$\max_{\substack{\{m(\tau(z),\zeta)\},\\ \{m(\tau(\zeta'),\zeta)\},\\ \tilde{\Pi}(\tau(z_0),\zeta),\\ \tilde{\Pi}(\tau(\zeta_0),\zeta)}} \Pi(\zeta) = \underbrace{\int_{\tau(z_0)} \left[\Big(\tau(z) - g^{-1}(\tau(z))\Big) m(\tau(z),\zeta)^{\frac{\sigma-1}{\sigma}} - c(\zeta)\, m(\tau(z),\zeta) - \tilde{\Pi}(\tau(z_0),\zeta)\right] d\tau(z) M_z}_{\text{profits from downstream firms}}$$
$$+ \underbrace{\int_{\tau(\zeta_0)} \left[\Big(\tau(\zeta') - g^{-1}(\tau(\zeta'))\Big) m(\tau(\zeta'),\zeta)^{\frac{\sigma-1}{\sigma}} - c(\zeta)\, m(\tau(\zeta'),\zeta) - \tilde{\Pi}(\tau(\zeta_0),\zeta)\right] d\tau(\zeta') M_\zeta}_{\text{profits from other upstream firms}}.$$

where:

$$\tau(\gamma) = A_\gamma \gamma^\theta \theta y(l(\gamma), m(\gamma))^{\theta-1} \left[\alpha l(\gamma)^{\frac{\eta-1}{\eta}} + (1-\alpha) m(\gamma)^{\frac{\eta-1}{\eta}}\right]^{\frac{1}{\eta-1}} (1-\alpha) m(\gamma)^{-\frac{1}{\eta}} \frac{\sigma}{\sigma-1} m(\gamma)^{\frac{1}{\sigma}}.$$

represents the buyer's heterogeneous valuation for the good sold by the seller, a strictly increasing function of respective productivities $\gamma = z, \zeta'$. $\tau(\gamma)$ is distributed according to a pareto distribution with tail parameter $\rho := \frac{\kappa\sigma}{\sigma-1}$[6]. Notice that $\rho > \sigma$, since $\kappa > \sigma - 1$. Moreover, $g(\tau)$ is the hazard function of the τ distribution, given by:

$$g(\tau) = \frac{g(\tau)}{1 - G(\tau)} = \frac{\rho}{\tau}$$

[6]In particular, under the guess, notice that the term:

$$A_\gamma \gamma^\theta \theta y(l(\gamma), m(\gamma))^{\theta-1} \left[\alpha l(\gamma)^{\frac{\eta-1}{\eta}} + (1-\alpha) m(\gamma)^{\frac{\eta-1}{\eta}}\right]^{\frac{1}{\eta-1}} (1-\alpha) m(\gamma)^{-\frac{1}{\eta}} = p_m$$

Since the term on the left of the equality is the marginal revenue product of the bundle of materials, thus equal to its price p_m, taken as given by the seller infinitesimal firm and independent of firm heterogeneity γ. Moreover, recall that

$$m(\gamma)^{\frac{1}{\sigma}} = \left(\frac{\gamma}{\gamma_0}\right)^{\frac{\sigma-1}{\sigma}} m(\gamma_0)^{\frac{1}{\sigma}} \Rightarrow \tau(\gamma) = \gamma^{\frac{\sigma-1}{\sigma}} p_m m(\gamma_0)^{\frac{1}{\sigma}}$$

Therefore, $\tau(\gamma)$ is distributed according to a pareto distribution with tail parameter $\rho := \frac{\kappa\sigma}{\sigma-1}$.

Let D_m denote the total demand for intermediates in the economy:

$$D_m := \int_{z_0}\int_{\zeta_0} m(z,\zeta')\mu(\zeta')M_\zeta\, d\zeta'\, \mu(z)M_z + \int_{\zeta_0}\int_{\zeta_0} m(\zeta,\zeta')\mu(\zeta')d\zeta' M_\zeta\, \mu(\zeta)M_\zeta.$$

Denote

$$s_y := \frac{p(\zeta_0)y(\zeta_0)}{D_m}$$

as the revenue share of upstream firms with the smallest upstream type ζ_0.

Then, we have the following proposition:

Proposition 1. *The B2B mechanism design equilibrium is characterized by the following solution:*

$$P(\gamma,\zeta) = p(\zeta)m(\gamma,\zeta) + F_\gamma\left(\frac{p(\zeta)}{p_m}\right)^{1-\sigma}, \tag{1}$$

$$p(\zeta) = \frac{\rho}{\rho-1}c(\zeta), \quad \rho = \frac{\sigma\kappa}{\sigma-1}, \quad p_m = \left(\int_{\zeta_0} p(\zeta)^{1-\sigma}M_\zeta\mu(\zeta)d\zeta\right)^{\frac{1}{1-\sigma}}, \quad \gamma \in \{z,\zeta\}. \tag{2}$$

The flat fee is given by:

$$F_\gamma(\zeta) = Revenue(\gamma_0)\frac{1}{\left(\frac{\alpha}{1-\alpha}\right)^\eta\left(\frac{p_m}{w}\right)^{\eta-1}+1}\left(\frac{1}{\sigma}\right)\left(\frac{p(\zeta)}{p_m}\right)^{1-\sigma} \quad \gamma \in \{z,\zeta\}. \tag{3}$$

Denote by F_z and F_ζ the flat fee charged by the average supplier $\tilde{\zeta}$. Then:

$$F_\gamma = A_\gamma\gamma_0^\theta y(l_0,m_0)^\theta\frac{1}{\left(\frac{\alpha}{1-\alpha}\right)^\eta\left(\frac{p_m}{w}\right)^{\eta-1}+1}\left(\frac{1}{\sigma}\right)\left(\frac{1}{M_\zeta}\right). \quad \gamma = z,\zeta \tag{4}$$

The revenue function is given by:

$$Revenue(\gamma) = A_\gamma q(\gamma)^\theta, \quad \gamma \in \{z,\zeta\}, \tag{5}$$

where

$$\theta = \frac{\sigma-1}{\sigma}, \quad A_z = Y^{\frac{1}{\sigma}}, \quad A_\zeta = \frac{D_m^{\frac{1-\sigma}{\sigma}}(p_mD_m + F_zM_z)}{1 - s_y\frac{1}{\left(\frac{\alpha}{1-\alpha}\right)^\eta\left(\frac{p_m}{w}\right)^{\eta-1}+1}}. \tag{6}$$

The input demand functions satisfy:

$$m(\gamma, \zeta') = \left(\frac{p(\zeta')}{p_m}\right)^{-\sigma} m(\gamma), \quad \gamma \in \{z, \zeta\}, \tag{7}$$

$$m(\gamma) = \left(\frac{\alpha}{1-\alpha}\frac{p_m}{w}\right)^{-\eta} l(\gamma), \quad \gamma \in \{z, \zeta\}. \tag{8}$$

The proof is provided in Appendix B.2. Notice that it is always optimal for the seller firm to serve all buyer types. While excluding the lowest type $\gamma_0, \quad \gamma = z, \zeta$ would allow the seller to charge a higher flat fee, doing so would result in the loss of demand associated with γ_0. Under the Pareto distribution, the mass of firms of type γ_0 is large enough that the seller never finds it optimal to exclude them[7]. As a result, including all types does not tighten the incentive constraints sufficiently to justify exclusion. In other words, there is no benefit in excluding any type. See the Appendix B.4 for more details.

Definition 2. *Decentralized Non-Linear Pricing Equilibrium*
A decentralized nonlinear pricing equilibrium is a collection of linear prices $\{p(\zeta), p(z)\}$, flat fees $\{F_z(\zeta), F_\zeta(\zeta)\}$, wage w, and quantities $\{y(z), q(z), l(z), m(z, \zeta'), l(\zeta), m(\zeta, \zeta'), M_z, M_\zeta\}$ such that, given technology $\{z, \zeta\}$:

i) The final good producer demands downstream goods to minimize costs.

ii) Both upstream and downstream producers demand intermediate inputs and labor to minimize costs.

iii) Each linear price equals marginal cost times the markup.

iv) The linear price markup is given by $\frac{\sigma}{\sigma-1}$ for downstream firms and $\frac{\rho}{\rho-1}$ for upstream firms.

v) Every firm pays a transfer $F_\gamma(\zeta)$, where $\gamma \in \{z, \zeta\}$, such that the lowest types ζ_0, z_0 have zero surplus from transacting with upstream sellers (i.e., the Individual Rationality constraint binds).

vi) Entrants earn zero expected profit.

[7] In the language of Mechanism Design, this can also be seen from the fact that the virtual utility under the Pareto distribution remains strictly positive, implying that even the lowest type contributes positively. Virtual utility represents the profits a seller obtains when accounting for both the direct gains from serving a firm type and the information rents required to prevent higher types from mimicking lower types. When virtual utility is positive, it is optimal to serve every firm type.

vii) Feasibility constraints hold with equality.

Lemma 2. ***Existence and Uniqueness*** *(see Appendix C.4 for proof)*
If $\eta = 1$, then both the linear and nonlinear pricing decentralized equilibria exist and are unique.

3.4 Equilibrium Characterization

In this section, we describe key features of the equilibrium.

3.4.1 Unit Prices

Figure 4 plots the unit price as a function of downstream firms' productivity (with the upstream buyer firm exhibiting a specular relationship), along with the unit price under the linear pricing equilibrium and the marginal cost. As described in Proposition 1, the total transfer follows a two-part tariff structure, consisting of a flat fee and a linear price component. The unit price paid by firm $\gamma = (z, \zeta')$ to firm ζ is given by:

$$p_{\text{unit}}(\gamma, \zeta) = \frac{P(\gamma, \zeta)}{m(\gamma, \zeta)} = \frac{F_\gamma(\zeta)}{m(\gamma, \zeta)} + p(\zeta), \quad \gamma = (z, \zeta').$$

The incidence of the flat fee on the unit price decreases as the quantity purchased increases, which is directly linked to buyer firm productivity. For low-productivity downstream firms, the flat fee constitutes the majority of the unit price, whereas for high-productivity firms, the unit price converges to the linear component of the two-part tariff.

High-productivity firms pay a lower unit price relative to a linear pricing setup, whereas low-productivity firms face a higher unit price. The total markup paid by each firm is given by the ratio of the total unit price to the marginal cost. Consequently, the total markup decreases with the quantity purchased, as the per-unit price declines.

Although the per-unit price varies with buyer firm productivity, the quantity allocation is entirely determined by the linear component of the two-part tariff. The allocation of purchased quantities is driven by the linear part of the tariff because the flat fee does not affect the marginal purchase decision (i.e., it does not influence the intensive margin), and it is set to ensure that every buyer firm participates in the transaction (i.e., it does not impact the extensive margin). The linear component of the two-part tariff features a homogeneous markup determined by $\rho = \frac{\kappa\sigma}{\sigma-1}$, which depends on the elasticity of substitution across varieties and the shape parameter of the Pareto distribution. We refer to this markup as the *allocative markup* to distinguish it from the total markup described above.

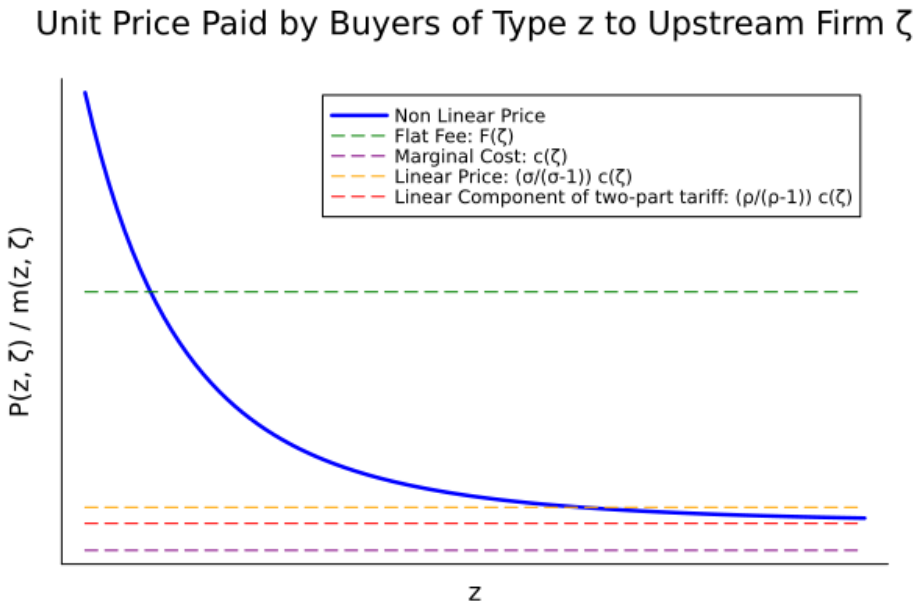


Figure 4: Unit Price Function

Notes: This figure illustrates the unit price function, where the pricing mechanism follows a nonlinear structure with reference levels. $F(\zeta) = 5$; $c(\zeta) = 2$; $\sigma = 3$; $\kappa = 2.74$; $m(z_0, \zeta) = 1$.

The remaining portion of the markup is referred to as the *redistributive markup*, which does not distort quantity allocations but instead redistributes rents from the buyer to the seller.

Notably, the allocative markup is lower than the total markup in the linear pricing case (where the total markup is allocative), as $\rho > \sigma \Leftrightarrow \kappa > \sigma - 1$, which is a necessary condition for finite output. [8] Using the terminology of Edmond et al. (2023), these allocative markups act as a uniform tax on production, reducing labor demand across all firms. As described in Lemma 1, conditional on entry, a Planner would aim to eliminate this uniform tax to restore marginal cost pricing. Consequently, since the allocative markup is lower under the non-linear pricing benchmark, the production decisions of existing firms align more closely with the Planner's optimal allocation, making the economy, conditional on entry, more efficient.

[8] Recall that in this economy, output is proportional to $Y \propto \tilde{z} = \left[\mathbb{E}_z z^{\sigma-1}\right]^{\frac{1}{\sigma-1}}$.

3.4.2 Firm Profits

While the flat fee does not distort the quantity purchased, it constitutes a transfer of profits from the buyer to the seller firm. Consequently, it affects firm entry decisions by distorting average profits. The profit functions under the linear pricing equilibrium are given by:

$$\Pi(z) = \frac{1}{\sigma - 1} c(z) \left(\frac{z}{\tilde{z}} \right)^{\sigma} M_z^{\frac{\sigma}{1-\sigma}} Y, \quad \text{for downstream firms,}$$
$$\Pi(\zeta) = \frac{1}{\sigma - 1} c(\zeta) \left(\frac{\zeta}{\tilde{\zeta}} \right)^{\sigma} M_\zeta^{\frac{\sigma}{1-\sigma}} D_m, \quad \text{for upstream firms.}$$

On the other hand, the profit functions under the non-linear pricing equilibrium are given by:

$$\Pi(z) = \frac{1}{\sigma - 1} c(z) \left(\frac{z}{\tilde{z}} \right)^{\sigma} M_z^{\frac{\sigma}{1-\sigma}} Y - F_z M_\zeta, \quad \text{for downstream firms,}$$

$$\Pi(\zeta) = \underbrace{\frac{1}{\rho - 1} c(\zeta) \left(\frac{\zeta}{\tilde{\zeta}} \right)^{\sigma} M_\zeta^{\frac{\sigma}{1-\sigma}} D_m}_{\text{Profits from the linear price component}} + \underbrace{\left(\frac{p(\zeta)}{p_m} \right)^{1-\sigma} M_\zeta (F_z M_z + F_\zeta M_\zeta)}_{\text{Collected flat fees}} - \underbrace{F_\zeta M_\zeta}_{\text{Flat fees paid}}, \quad \text{for upstream firms.}$$

Notably, for the upstream firm, in expectation, the flat fee collected from and paid to other upstream firms cancels out exactly[9]. Profits are distorted due to a redistribution of resources across firms. Figure 5 presents a comparison of profit functions under linear and nonlinear equilibrium concepts for both downstream and upstream firms in partial equilibrium (i.e., holding aggregate quantities and prices constant).

For downstream firms, the difference between the two profit functions manifests as a downward shift of the curve, since their variable profits remain unchanged while they pay a flat fee to upstream firms. For upstream firms, the distinction is more nuanced.

[9] By definition $F_z = \mathbb{E}_\zeta [F_z(\zeta)]$ and $F_\zeta = \mathbb{E}_\zeta [F_\zeta(\zeta)]$. Thus:

$$\mathbb{E}[\Pi(\zeta)] = \underbrace{\frac{1}{\rho - 1} c(\zeta) M_\zeta^{\frac{\sigma}{1-\sigma}} D_m}_{\text{Profits from the linear price component}} + \underbrace{(F_z M_z + F_\zeta M_\zeta)}_{\text{Collected flat fees}} - \underbrace{F_\zeta M_\zeta}_{\text{Flat fees paid}}, \quad \text{for upstream firms.}$$

First, the profits derived from the linear component of their two-part tariff are lower, as they charge a reduced markup. Additionally, they appropriate the average flat fees paid by all firms in proportion to their productivity while also bearing the aggregate flat fee themselves.

This implies that the wedge between profits in the linear and nonlinear pricing equilibria increases with firm productivity, with the most productive firms benefiting the most from the nonlinear pricing equilibrium. For low-productivity upstream firms, conditional profits remain closer to those under the linear pricing equilibrium and may even be lower (though not necessarily), as depicted in the figure.

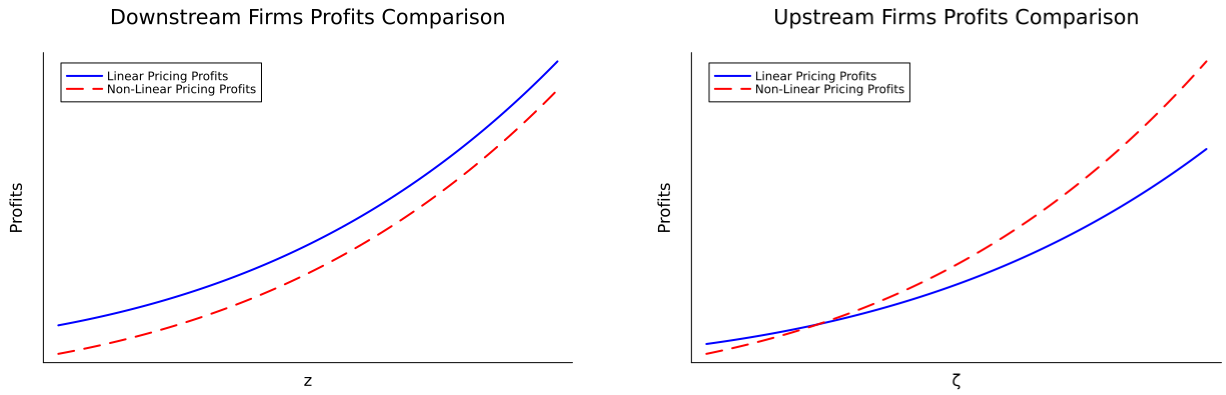


Figure 5: Profit Comparison Linear vs Nonlinear equilibrium

Lastly, we highlight that the profits of the lowest-type firm are not zero as in the classic textbook Mechanism Design Problem. To see this, consider a downstream firm of type z_0. It is well known that in a CES demand setting under linear prices, variable profits constitute a fraction $\frac{1}{\sigma}$ of revenue, while variable costs take up a fraction $\frac{\sigma-1}{\sigma}$ of revenue. Denote s_m as the intermediate input share of total variable costs. In the appendix, we show that the profit of the lowest type z_0 is given by:

$$\Pi(z_0) = \text{Revenue}(z_0)\left[1 - \frac{1}{\sigma}s_m - \frac{\sigma-1}{\sigma}\right].$$

Notably, when $s_m = 0$, profits revert to a share $\frac{1}{\sigma}$ of total revenue as in the linear benchmark, while as the intermediate input share s_m approaches one, total profit converges to zero. This result indicates that the lowest-type buyer receives some rent due to their ability to substitute intermediates with labor.

3.4.3 Inefficiencies

In this economy, deviations from the efficient allocation are fully characterized by four key variables (see Appendix C for details):

$$\tau := \frac{l_z}{l_{z,eff}}, \quad q\left(\frac{w}{p_m}\right) := \frac{l_\zeta M_\zeta}{L}, \quad \theta\left(\frac{w}{p_m}\right) := \frac{l_\zeta}{l_z}, \quad K\left(\frac{w}{p_m}\right) := \frac{L}{F\Phi\left(\frac{w}{p_m}\right)}$$

where l_z and l_ζ denote the labor used by the average downstream and upstream firms, respectively. The term $l_{z,eff}$ represents the labor used by the average downstream firm under the efficient allocation. L denotes the total labor used in production, F is a constant common to all equilibrium notions, and $\Phi\left(\frac{w}{p_m}\right)$ is a function of $\frac{w}{p_m}$ that remains consistent across equilibrium definitions (see Appendix C for further details):

$$F := (\tilde{z})^{1-\sigma}\,\sigma(1-\delta)c_e\,(1-\alpha)^{-1}\,\alpha\Theta^{(\sigma-1)},$$
$$\Phi\left(\frac{w}{p_m}\right) := \left(\frac{w}{p_m}\right)^{(\sigma-1)\alpha}.$$

The variable τ measures the deviation of labor used by the average downstream firm from its efficient level. The function q represents the share of production labor allocated to upstream firms, while θ captures the relative labor usage between the average downstream and upstream firms. The function K acts as a scaling factor that adjusts total labor used in production relative to the function $\Phi(w)$.

Lemma 3. *Under the efficient allocation, the equilibrium conditions satisfy:*

$$\tau = 1, \quad q = s_m, \quad K = 1, \quad \theta = 1.$$

Thus, if $\tau \neq 1$, the labor used by the average downstream firm is distorted, indicating that it employs either too much or too little labor. If $\theta \neq 1$, labor allocation between upstream and downstream firms is inefficient, making θ a measure of misallocation across these two sectors. If $q \neq s_m$, the aggregate allocation of labor between upstream and downstream sectors is inefficient. Together with θ, q serves as an indicator of misallocation in sectoral entry. Lastly, if $K \neq 1$, total labor in production does not scale appropriately with wages, indicating that labor allocation in production is inefficient relative to labor used in entry.

Proposition 2. *Let* $s := \left(\frac{z_0}{\tilde{z}}\right)^{\sigma-1}$. *Under the nonlinear pricing equilibrium, the economy is distorted as follows:*

$$\tau_{nlp} = \frac{\sigma-1}{\sigma}\frac{1}{1-ss_m\left(\frac{w}{p_m}\right)}, \quad q_{nlp} = \frac{\rho-1}{\rho}s_m\left(\frac{w}{p_m}\right), \quad \theta_{nlp} = \frac{1}{\sigma-1}(1-ss_m\left(\frac{w}{p_m}\right))\left[\frac{1}{\rho-1}+\frac{1-q_{nlp}}{q_{nlp}}\frac{ss_m\left(\frac{w}{p_m}\right)}{\sigma-1}\right]^{-1},$$

$$K_{nlp} = \left(\frac{\rho-1}{\rho}\right)^{1-\sigma}\frac{\rho}{\rho-1}\frac{1}{\sigma}\left[\frac{1}{\rho-1}+\frac{1-q_{nlp}}{q_{nlp}}\frac{ss_m\left(\frac{w}{p_m}\right)}{\sigma-1}\right]^{-1}.$$

On the other hand, under linear pricing:

$$\tau_{lp} = \frac{\sigma-1}{\sigma}, \quad q_{lp} = \frac{\sigma-1}{\sigma}s_m\left(\frac{w}{p_m}\right), \quad \theta_{lp} = 1, \quad K_{lp} = \left(\frac{\sigma}{\sigma-1}\right)^{\sigma-1}.$$

Proposition 2 highlights the wedges through which the inefficiencies generated by each equilibrium notion distort the equilibrium. According to Lemma 1, the efficient benchmark is achieved through a decentralized Linear Pricing equilibrium with an output subsidy that restores marginal cost pricing, conditional on entry. In the Linear Pricing equilibrium, the constant markup charged by firms acts as a tax on production, thereby reducing firm output and labor demand. Moreover, since the economy features linkages between firms through intermediate goods, the effect of this output tax is amplified by a multiplier, in the spirit of Jones (2011). To restore the efficient equilibrium, the government requires a single policy instrument: an output subsidy financed by a lump-sum tax on the final consumer (see Lemma 1).

The nonlinear pricing equilibrium introduces three types of inefficiencies: (i) allocative markups that function as an output tax on production, (ii) a redistribution of rents from downstream to upstream firms that effectively acts as an entry tax, and (iii) an inefficient markup charged by upstream firms, which again distorts entry. The combination of these inefficiencies distorts equilibrium allocations. First, firms charge a markup conditional on entry, distorting input choices. In particular, input decisions by downstream firms are affected by the standard CES markup $\mu = \frac{\sigma}{\sigma-1}$. For upstream firms, input choices are distorted by a lower allocative markup—specifically, the markup applied to the linear component of the two-part tariff. This markup is lower than the standard CES markup under linear pricing because $\rho > \sigma$, bringing the economy closer to the efficient allocation in this regard. However, since the allocative markup is now heterogeneous between

the downstream and upstream sectors, resource misallocation arises across the two sectors, as is common in models with heterogeneous markups. To correct this inefficiency, the planner would require the same policy instrument as in the linear pricing case—an output subsidy—but now differentiated across sectors to restore marginal cost pricing, conditional on entry, for the linear component of the two-part tariff.

While firm production becomes more efficient, the seller firm's surplus extraction introduces a second inefficiency that acts as an entry tax. Efficiency requires that entry occurs until entry costs (and thus, in equilibrium, expected profits) are equal to the expected marginal value of entry—the sum of producer and consumer surplus. By Lemma 1, this condition is satisfied if firms earn the profits associated with a linear pricing scheme with a markup equal to the CES markup $\frac{\sigma}{\sigma-1}$. As illustrated in Figure 5, downstream firms pay a flat fee to upstream firms, reducing their expected profits. While upstream firms also pay flat fees to other upstream firms, these payments cancel out in expectation, as explained earlier. Therefore, this inefficiency manifests as a redistribution of resources from downstream to upstream firms, effectively functioning as an entry tax collected by upstream firms at the expense of downstream firms.[10] A planner would eliminate this inefficiency by implementing a lump-sum tax on upstream firms, rebated to downstream firms, exactly equal to the surplus extracted.

The third inefficiency arises because the profits associated with the linear component of the two-part tariff for upstream firms are no longer determined by the CES markup but by a lower one. Consequently, entry is distorted, as expected profits do not equate the marginal value of entry (the love of variety effect). In this case, a planner would remove the inefficiency by providing a transfer to upstream firms, financed by a lump-sum tax on households, equal to the difference between the variable profits under linear pricing and the profits from the linear component of the two-part tariff in the nonlinear pricing equilibrium.

Lemma 4. ***Optimal Policy in the Nonlinear Pricing Equilibrium***
The efficient benchmark is achieved in a decentralized Nonlinear Pricing Equilibrium with a planner implementing the following policies:

- *An output subsidy of $\tau_z = \frac{\sigma-1}{\sigma}$ for downstream firms and an output subsidy of $\tau_\zeta = \frac{\rho-1}{\rho}$ for upstream firms. Both subsidies can be financed via a lump-sum tax on the final consumer.*

[10] This entry tax would be absent in a purely horizontal economy—i.e., one without vertical linkages—where firms sell to both other firms and consumers without an upstream-downstream structure.

- *A lump-sum entry subsidy to downstream firms equal to* $\frac{1}{1-\delta}F_z M_\zeta$*, paid by upstream firms, which in turn pay a lump-sum entry tax equal to* $\frac{1}{1-\delta}F_z M_z$*.*
- *A lump-sum transfer to upstream firms equal to* $\tau_\zeta c(\tilde{\zeta}) D_\zeta M_\zeta^{\frac{\sigma}{1-\sigma}} \left(\frac{1}{\sigma-1} - \frac{1}{\rho-1}\right)$*, financed via a lump-sum tax on the final consumer.*

The proof follows directly from the discussion in this section, recognizing that once these policy instruments are implemented, the resulting equilibrium coincides with the one characterized in Lemma 4.

While the output tax inefficiency is mitigated under the nonlinear pricing scheme, the entry tax introduces an additional distortion that could make the economy more inefficient. Determining which of these effects dominates is ultimately a quantitative question, which we address in the following sections of the paper.

4 Model Calibration and Quantification

This section outlines our approach to calibrating the model parameters and quantifying the welfare implications of linear versus nonlinear pricing relative to efficient pricing. We first describe the calibration strategy and parameter values, followed by an analysis of the quantitative results in relation to the three economic distortions identified in Section 3.4.3.

4.1 Preliminary Calibration

To ensure equilibrium existence and uniqueness, we set $\eta = 1$, which implies Cobb-Douglas production functions throughout the model. This assumption enhances analytical tractability while preserving economic relevance. The remaining parameters are calibrated to match key empirical moments from firm-level data.

Table 6 presents the parameter values used in our quantitative analysis. The labor share in production ($\alpha = 0.38$) is calibrated directly from input-output data, reflecting the proportion of total costs attributable to labor inputs. We provide sensitivity analysis with respect to two values for the elasticity of substitution for the material bundle. In particular, we set $\sigma = 5$ or $\sigma = 3$, aligning with estimates in the literature on input substitutability and misallocation (Hsieh and Klenow (2009)). This choice implies a linear markup charged by downstream firms of respectively $\mu = 1.5$ or $\mu = 1.25$. The firm exit

rate ($\delta = 0.15$) is calibrated using business dynamics statistics, representing the annual probability of firm exit. The entry cost ($c_e = 0.16$) is drawn from Hopenhayn et al. (2022). The Pareto tail parameter (κ) governs the productivity distribution and is calibrated to match the observed distribution of labor. In particular the data reveal that firm size distribution measured as total number of workers follows a Pareto distribution with Tail $\nu = 1.48$[11]. The model-implied firm size distribution, measured as total labor, follows a Pareto distribution with a tail parameter of $\frac{\kappa}{\sigma-1}$. This arises from the labor allocation function $l(\gamma) = l(\tilde{\gamma})\left(\frac{\gamma}{\tilde{\gamma}}\right)^{\sigma-1}$. Thus, given the assumed σ and the estimated labor distribution tail of ≈ 1.48, we calibrate κ accordingly.

Table 6: Model Parameters

	Value	Source
Labor share in production (α)	0.38	Calibrated from data
Material bundle elasticity (σ)	5 or 3	Assumed
Firm exit rate (δ)	0.15	Calibrated from data
Entry cost (c_e)	0.16	Assumed
Pareto tail (κ)	5.92 or 2.96	Calibrated from data

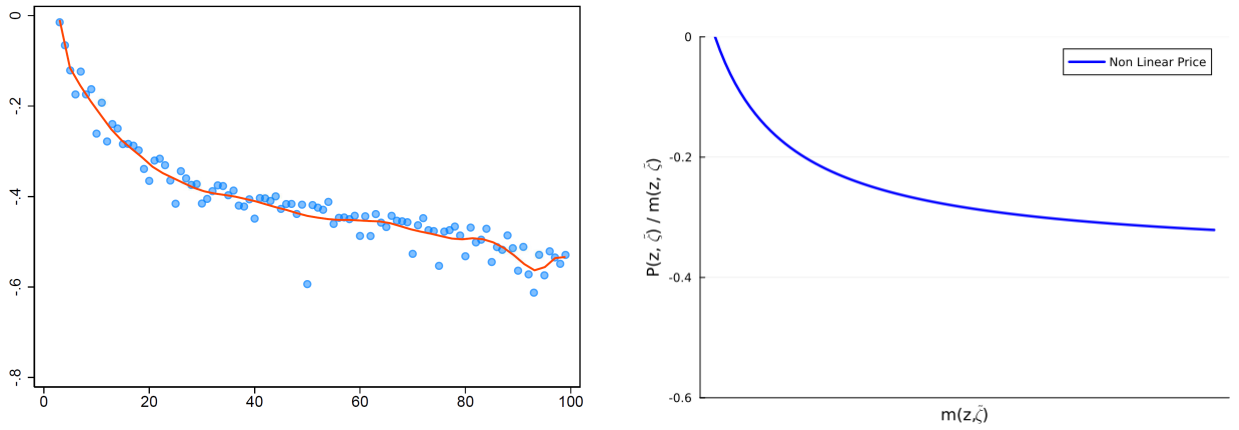
Once calibrated, we assess the model's empirical validity by comparing the equilibrium nonlinear unit prices with those observed in the data. Figure 6 presents a comparison, with the left panel showing empirically estimated nonlinear prices and the right panel displaying the model-implied equilibrium prices. In particular, the right panel illustrates the unit prices charged by the average upstream firm $\tilde{\zeta}$ to downstream firms as a function of the quantity purchased, corresponding to values of z ranging from z_0 to the 95th percentile of the z distribution, when the model is calibrated with $\sigma = 5$.

It is important to note that the nonlinear pricing pattern observed in the data is not explicitly targeted by the model. Instead, the model is calibrated using firm size distribution statistics and parameter estimates from the literature. Nevertheless, the nonlinear pricing model successfully reproduces the shape of the negative relationship between unit prices and purchase quantities, a pattern observed in the data, where larger purchases receive substantial quantity discounts. Furthermore, the model captures the economic significance of these quantity discounts, which reach approximately 40% in both the model and

[11]This is the Tail estimate using an MLE approach with firms with at least two employees.

the data[12].

Figure 6: Calibrated Model Unit Prices



(a) Nonlinear prices in data

(b) Model-implied nonlinear prices

Note: The left panel shows nonlinear prices observed in the data, while the right panel presents the corresponding equilibrium prices implied by the calibrated model with $\sigma = 5$. These prices represent the unit prices charged by the average upstream firm $\tilde{\zeta}$ to downstream buyers as a function of purchase quantities.

4.2 Quantification

We now quantify the economic implications of linear versus nonlinear pricing. This analysis proceeds through three counterfactual exercises. First, we compare both linear and nonlinear pricing regimes against the efficient pricing benchmark. Second, to highlight the different channels through which nonlinear pricing influences outcomes, we introduce and solve for the equilibrium under three distinct policy interventions, implemented one at a time.

First, we analyze an equilibrium with a lump-sum transfer that offsets the extraction of resources from upstream firms at the expense of downstream firms, as outlined in bullet point 2 of Lemma 4. Second, we introduce an output subsidy that restores marginal cost pricing conditional on entry, i.e.,

$$\tau_z = \frac{\sigma - 1}{\sigma}, \quad \tau_\zeta = \frac{\rho - 1}{\rho}.$$

[12]The figure with $\sigma = 3$ is very similar with quantity discounts reaching 25% in value.

Third, we solve for an equilibrium in which the government implements a uniform output subsidy as if the economy was linear applied to both upstream and downstream firms i.e.,

$$\tau_z = \frac{\sigma - 1}{\sigma}, \quad \tau_\zeta = \frac{\sigma - 1}{\sigma}.$$

Notably, the equilibrium featuring both transfers (bullet points 2 and 3 in Lemma 4) is equivalent to the linear pricing equilibrium, while an equilibrium incorporating the full optimal policy restores efficiency.

Nonlinear Equilibrium, Linear Equilibrium, and Efficient Equilibrium

Table 7 presents the ratios of key variables under both pricing regimes relative to their efficient levels.

Table 7: Equilibrium Outcomes Relative to Efficiency

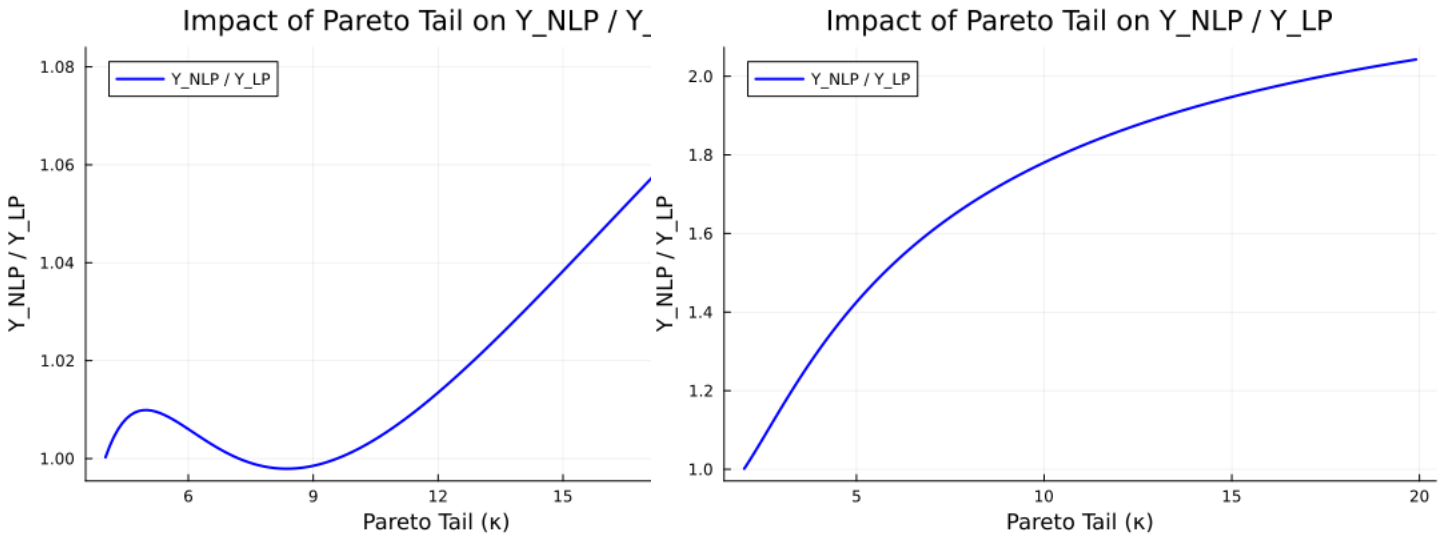
	$\sigma = 3$		$\sigma = 5$	
Variable	LP / Eff	NLP / Eff	LP / Eff	NLP / Eff
Total Firms (M)	1.2159	1.0967	1.1508	0.9991
Upstream Mass of Firms (M_ζ)	0.8106	0.6934	0.9206	0.6824
Downstream Mass of Firms (M_z)	1.8772	1.7548	1.5263	1.5158
Welfare (Y)	0.4411	0.5075	0.7538	0.7586
Average Firm Output (y)	0.2899	0.3489	0.5374	0.5439
$Y_{\text{NLP}}/Y_{\text{LP}}$	1.1506		1.0065	

The quantitative results indicate that the nonlinear pricing equilibrium is closer to the efficient benchmark compared to the linear pricing equilibrium. Entry (i.e., total mass of firms) under nonlinear pricing is lower by approximately 10–15% relative to linear pricing, depending on the value of σ. Conversely, the average output per downstream firm is 1–20% higher than in the linear pricing equilibrium, depending on σ. The cumulative effect of these two forces results in an overall efficiency gain of 15% when $\sigma = 3$ and 0.6% when $\sigma = 5$. Nevertheless, both pricing equilibria remain significantly below the efficient benchmark, with GDP approximately 50% lower when $\sigma = 3$ and 25% lower when $\sigma = 5$, reflecting excessive entry and reduced firm-level output.

What determines differences in aggregate variables across linear and nonlinear pricing equilibria? First, when the elasticity of substitution (σ) is high, reducing the "love of variety" effect, the linear pricing economy converges toward efficiency, as firms' input distortions diminish. Moreover, price indices become less responsive to distortions, making the nonlinear pricing equilibrium resemble the linear pricing equilibrium. Second, the difference between the two equilibrium notions depends on how close κ is to $\sigma - 1$. In a purely horizontal economy, as $\kappa \to \sigma - 1$, the linear pricing benchmark becomes isomorphic to the nonlinear pricing benchmark in terms of aggregate variables, since the allocative markup converges to the linear markup and transfers cancel out for the average firm. In the two-layered structure considered here, there remains a redistribution effect from downstream to upstream firms, but its quantitative impact is relatively minor.

Figure 7 illustrates this concept by plotting GDP under nonlinear pricing relative to linear pricing for different values of κ. As κ approaches $\sigma - 1$, GDP under nonlinear pricing converges to that under linear pricing. Additionally, GDP may even be lower, as shown by the minimum in the left panel, where the GDP ratio falls below 1. Finally, nonlinear pricing significantly enhances aggregate efficiency for larger values of κ and smaller values of σ. For instance, when $\kappa = \sigma$, the nonlinear pricing equilibrium exhibits an efficiency gain of approximately 10% over the linear pricing equilibrium for $\sigma = 3$ but only 0.1% for $\sigma = 5$.

Figure 7: GDP Under Nonlinear Pricing Relative to Linear Pricing



(a) $\sigma = 5$ (b) $\sigma = 3$

Note: Both panels illustrate GDP under the nonlinear pricing equilibrium relative to the linear pricing equilibrium ($\frac{Y_{NLP}}{Y_{LP}}$)for different values of the Pareto tail parameter κ. The left panel corresponds to $\sigma = 5$, while the right panel corresponds to $\sigma = 3$.

Introducing Policy

Next, we introduce one policy instrument at a time, as outlined in Lemma 4. Table 8 reports the results of these policy experiments for the two assumed values of σ. First, we introduce a transfer from upstream to downstream firms. Under the current parameterization, the transfer has a negligible impact on aggregate variables for large values of σ, leaving outcomes largely unchanged. In contrast, when $\sigma = 3$, the transfer reduces aggregate GDP by lowering entry in the upstream sector.

We then consider a differentiated subsidy for downstream and upstream firms (τ_z and τ_ζ), designed to restore marginal cost pricing conditional on entry. This policy increases aggregate GDP by approximately 50% ($\sigma = 3$) or 23% ($\sigma = 5$) relative to the realized nonlinear pricing outcome, primarily by increasing output at the average downstream firm $\tilde{z}$, while also reducing firm entry.

The largest welfare gains arise from a policy that provides a uniform output subsidy to both upstream and downstream firms. Specifically, aggregate GDP under this policy increases by approximately 70% ($\sigma = 3$) or 27% ($\sigma = 5$) relative to the realized nonlinear pricing outcome, driven by higher average output per downstream firm.

Under this policy, aggregate GDP aligns closely with the efficient benchmark, suggesting that—quantitatively—the government can restore efficiency through a simple output subsidy. This mechanism functions as if the economy were a linear economy, effectively correcting the primary inefficiencies identified in Lemma 4. Thus, a constant output subsidy applied to both upstream and downstream firms successfully addresses both input distortions and profit distortions.

Table 8: Policy Experiment Outcomes Relative to NLP Benchmark

	$\sigma = 3$					$\sigma = 5$				
Policy	M/M_{NLP}	$M_z/M_{z,\text{NLP}}$	$M_\zeta/M_{\zeta,\text{NLP}}$	y/y_{NLP}	Y/Y_{NLP}	M/M_{NLP}	$M_z/M_{z,\text{NLP}}$	$M_\zeta/M_{\zeta,\text{NLP}}$	y/y_{NLP}	Y/Y_{NLP}
Transfer	1.0000	1.0698	0.8471	0.8164	0.8540	1.0000	1.0069	0.9906	0.9893	0.9948
Subsidy (τ_z, τ_ζ)	0.7855	0.6436	1.0965	2.0492	1.5275	0.8592	0.7040	1.0705	1.6288	1.2300
Subsidy (τ_z, τ_z)	0.6207	0.3699	1.1705	3.3051	1.7031	0.8204	0.6224	1.0899	1.8633	1.2752

5 Conclusion

This paper investigates the macroeconomic implications of price discrimination in firm-to-firm transactions, focusing on its effects on firm entry, resource allocation, and aggregate welfare. Using detailed firm-to-firm transaction data, we document the widespread prevalence of price discrimination in supply chains.

To quantify the welfare effects of price discrimination, we develop a supply chain general equilibrium model in which firms can price discriminate. The model highlights how second- and third-degree price discrimination jointly determine firm-level production decisions. Relative to linear price setups, we find that while price discrimination mitigates inefficiencies in firm-level production, it reduces firm entry by distorting firm-level profit distribution. Quantitatively, these distortions result in aggregate welfare gains compared to an equilibrium where firms are constrained to offering linear prices, suggesting that linear price models overestimate the welfare costs of market power.

We characterize optimal policies that correct these distortions, restoring efficiency in a decentralized price discrimination setup equilibrium. Specifically, we show that a uniform output subsidy effectively restores efficiency by addressing output and firm entry distortions. In contrast, targeted transfers and differentiated subsidies that restore marginal cost pricing conditional on entry yield limited gains, particularly in economies with high substitution elasticities.

Our results underscore the importance of incorporating realistic pricing mechanisms into macroeconomic models to accurately measure the welfare effects of market power in supply chains.

References

Baqaee, D. and Farhi, E. (2020). Entry vs. rents: Aggregation with economies of scale. Technical report, National Bureau of Economic Research.

Berger, D., Herkenhoff, K., and Mongey, S. (2022). Labor market power. *American Economic Review*, 112(4):1147–1193.

Boehm, J., South, R., Oberfield, E., and Waseem, M. (2024). The network origins of firm dynamics: Contracting frictions and dynamism with long-term relationships.

Bornstein, G. and Peter, A. (2024). Nonlinear pricing and misallocation. Technical report, National Bureau of Economic Research.

Burstein, A., Cravino, J., and Rojas, M. (2024). Input price dispersion across buyers and misallocation. Technical report, Central Bank of Chile.

Cowan, S. (2016). Welfare-increasing third-degree price discrimination. *The RAND Journal of Economics*, 47(2):326–340.

De Loecker, J., Eeckhout, J., and Mongey, S. (2021). Quantifying market power and business dynamism in the macroeconomy. Technical report, National Bureau of Economic Research.

Edmond, C., Midrigan, V., and Xu, D. Y. (2023). How costly are markups? *Journal of Political Economy*, 131(7):1619–1675.

Fajgelbaum, P., Grossman, G. M., and Helpman, E. (2011). Income distribution, product quality, and international trade. *Journal of political Economy*, 119(4):721–765.

Hopenhayn, H., Neira, J., and Singhania, R. (2022). From population growth to firm demographics: Implications for concentration, entrepreneurship and the labor share. *Econometrica*, 90(4):1879–1914.

Hopenhayn, H. A. (1992). Entry, exit, and firm dynamics in long run equilibrium. *Econometrica: Journal of the Econometric Society*, pages 1127–1150.

Hsieh, C.-T. and Klenow, P. J. (2009). Misallocation and manufacturing tfp in china and india. *The Quarterly journal of economics*, 124(4):1403–1448.

Hsieh, C.-T. and Klenow, P. J. (2014). The life cycle of plants in india and mexico. *The Quarterly Journal of Economics*, 129(3):1035–1084.

Jones, C. I. (2011). Intermediate goods and weak links in the theory of economic development. *American Economic Journal: Macroeconomics*, 3(2):1–28.

Melitz, M. J. (2003). The impact of trade on intra-industry reallocations and aggregate industry productivity. *econometrica*, 71(6):1695–1725.

Appendix

A Quantity Discount evidence by industry

$$\ln\ p_{ijgm} = \beta_0 + \beta_1 \ln\ _{ijg} + \ln\ \psi_{ijg} + \ln\ \epsilon_{ijgm}$$

Sector	β_1	se	N obs (millions)	Adjusted R2
Agriculture	-0.480	(0.0002)	8.2	0.83
Mining	-0.375	(0.0008)	0.8	0.84
Manufacturing	-0.161	(0.0003)	313.6	0.92
Utilities	-0.174	(0.0005)	18.9	0.86
Construction	-0.722	(0.0006)	3.7	0.82
Retail and Wholesale	-0.181	(0.0002)	766.1	0.90
Transport and ICTs	-0.489	(0.0003)	39.3	0.85
Financial Services	-0.492	(0.0004)	15.5	0.79
Real Estate Services	-0.773	(0.0014)	1.7	0.81
Business Services	-0.562	(0.0003)	16.9	0.96
Personal Services	-0.498	(0.0010)	2.9	0.74

B Proofs

B.1 Proof of lemma 1

Proof. Considering the Kuhn-Tucker conditions for the social planner's problem inspired in Baqaee and Farhi (2020) theorem 1 proof.

Step 1: Formulating the Social Planner's Problem

Let the social planner maximize the utility function of a representative consumer, who derives utility from final consumption:

$$L = U(C) + \sum_{i \in N} \rho_i \left(Y_i - \sum_{k \in N} M_k x_{ki} - C_i \right) + \sum_{i \in N} \lambda_i \left(M_i F_i \left(\frac{y_i}{Y_i} \right) - 1 \right).$$

where $U(C)$ is the representative consumer's utility from total consumption, Y_i represents firm-level output, x_{ki} is the input from firm k used in production by firm k and M_i is the mass of firms of type i. ρ_i and λ_i are the Lagrange multiplier associated with the resource constraint for firm i and firm scale F_i

Firms trade intermediate goods with each other and charge marginal costs to consumers but apply markups to other firms.

Step 2: First-Order Conditions
aking the first-order conditions with respect to consumption, output, and factor inputs, we obtain:

$$\begin{aligned}
\frac{\partial L}{\partial C} &= \frac{\partial U}{\partial C} - \sum_{i \in N} \rho_i = 0, \\
\frac{\partial L}{\partial Y_i} &= \rho_i - \lambda_i M_i F_i' \left(\frac{y_i}{Y_i} \right) \frac{y_i}{Y_i} = 0, \quad \forall i, \\
\frac{\partial L}{\partial x_{ij}} &= -\rho_j M_i + \alpha_i A_i \frac{\partial f_i}{\partial x_{ij}} = 0, \quad \forall i, j.
\end{aligned}$$

Step 3: Equilibrium Conditions
The decentralized equilibrium consists of firms setting prices as follows:

$$\begin{aligned}
P_i &= MC_i, && \text{if selling to consumers,} \\
P_i &= \mu_i MC_i, && \text{if selling to other firms,}
\end{aligned}$$

where $\mu_i > 1$ represents the markup applied to firm-to-firm transactions. Market-clearing conditions require:

$$\begin{aligned} P_i Y_i &= \rho_i P_C, \\ MC_i A_i \frac{\partial f_i}{\partial x_{ij}} &= P_j, \\ \sum_{i \in N} \varepsilon(j,i) M_{E,j} &= M_i. \end{aligned}$$

Where $\varepsilon(j,i)$ represents the fraction of entering firms of type j that transition into firms of type i, A_i is the productivity parameter of firm i, P_C is the aggregate price index, F_i is the production function of firm i

Step 4: Simplification of First-Order Conditions

Using these equilibrium conditions, we rewrite the first-order conditions in a simplified form that characterizes the optimal allocation of resources and the zero-profit conditions for entry:

$$\begin{aligned} \rho_i &= \frac{P_i Y_i}{P_C}, \quad \forall i, \\ MC_i &= \frac{P_j}{A_i \frac{\partial f_i}{\partial x_{ij}}}, \quad \forall i, j, \\ M_i &= \sum_{j \in E} \varepsilon(j,i) M_{E,j}, \quad \forall i. \end{aligned}$$

These equations describe (i) firms pricing at marginal cost when selling to consumers, (ii) input demand following cost minimization, and (iii) the free-entry condition ensuring expected profits equal entry costs.

Step 5: Comparing Social Planner and Decentralized Equilibrium

By substituting these equilibrium conditions into the social planner's optimization problem, we compare it with the decentralized outcome. The planner's allocation satisfies:

$$\begin{aligned} P_i Y_i &= \rho_i P_C, \\ MC_i &= \frac{P_j}{A_i \frac{\partial f_i}{\partial x_{ij}}}, \\ \sum_{i \in N} \varepsilon(j,i) M_{E,j} &= M_i. \end{aligned}$$

In the decentralized economy, firms set prices as:

$$P_i = \mu_i MC_i, \quad \forall i,$$

for firm-to-firm trade. Comparing these conditions, we see that in the decentralized outcome, markups distort efficiency by increasing input prices for other firms, reducing output relative to the social planner's allocation. To restore efficiency, a subsidy that offsets markups in firm-to-firm trade is needed:

$$\tau_i = \frac{\mu_i - 1}{\mu_i}.$$

This subsidy eliminates markup distortions in firm-to-firm pricing while keeping marginal-cost pricing for final consumers.

By explicitly comparing the social planner's allocation with the decentralized equilibrium, we show that firms imposing markups on firm-to-firm trade distorts efficiency. A subsidy equal to $\tau_i = \frac{\mu_i - 1}{\mu_i}$ can correct these distortions, restoring optimal resource allocation.

□

B.2 Proof of Proposition 1

Assumption 2. Firm types follow a Pareto distribution with tail parameter κ.

Assumption 3. Firms are infinitesimally small.

By "infinitesimally small," we mean that the seller firm internalizes only first-order effects when solving its problem, as second-order effects vanish due to the infinitesimal nature of the firms.

We proceed by applying the method of guessing and verifying.

Guess 1: Total transfer takes the form of a two-part tariff:

$$P(\gamma, \zeta) = p(\zeta) m(\gamma, \zeta) + F_\gamma(\zeta), \quad \text{for} \quad \gamma = z \text{ or } \zeta.$$

Guess 2: The revenue functions are described as follows:

$$R(z) = z^\theta y\Big(l(z), m(z)\Big)^\theta A_z,$$

and

$$R(\zeta) = \zeta^{\theta} y\Big(l(\zeta), m(\zeta)\Big)^{\theta} A_{\zeta}.$$

B.3 Problem of the Seller

Solving the problem of the seller allows us to verify the initial guess. Note that firms purchasing inputs are also paying for labor.

In particular, define the buyer's profit function as:

$$\Pi(\gamma) = \max_{l(\gamma), \{m(\gamma,\zeta)\}} \gamma^{\theta} y\Big(l(\gamma), m(\gamma)\Big)^{\theta} A_{\gamma} - w l(\gamma) - \int_{\zeta_0} P(\gamma, \zeta)\mu(\zeta) M_{\zeta}\, d\zeta, \quad \gamma = z \text{ or } \zeta,$$

where this problem implicitly defines the buyer's demand for labor and the buyer's demand for materials as functions of the marginal product of each input, evaluated at the buyer type γ.

Index the infinitesimal firm we are analyzing by i, and let ζ denote its productivity. Under the guess, the i-th firm's marginal cost is constant and given by:

$$c(\zeta) = \frac{1}{\zeta} c,$$

where c is a function of prices. In particular, under the guess of the two-part tariff, the common component of the marginal cost is:

$$c(p_m, w) = \left(\alpha^{\eta} w^{1-\eta} + (1-\alpha) p_m^{1-\eta}\right)^{\frac{1}{1-\eta}},$$

where p_m is the CES index of the linear price of the two-part tariff:

$$p_m = \left(\int p_m(\zeta)^{1-\sigma} M_{\zeta} \mu(\zeta)\, d\zeta\right)^{\frac{1}{1-\sigma}}.$$

Define $\tilde{\Pi}(\gamma, \zeta) = \frac{d\Pi(\gamma)}{d\mu(\zeta) M_{\zeta}}, \quad \gamma = z \text{ or } \zeta$, as the *total buyer γ surplus from transacting with seller ζ*, i.e., the additional profits that the buyer gains from purchasing from the infinites-

imal seller i. Using the envelope theorem, this can be expressed as:

$$\tilde{\Pi}(\gamma,\zeta) = \gamma^{\theta}\theta y(l,m)^{\theta-1}\frac{\partial y(l,m)}{\partial m}\frac{\partial m}{\partial \mu(\zeta)M_{\zeta}} - P(\gamma,\zeta)$$

Using the CES assumption (suppressing indexes of aggregate variables):

$$\tilde{\Pi}(\gamma,\zeta) =$$

$$A_{\gamma}\gamma^{\theta}\theta y(l(\gamma),m(\gamma))^{\theta-1}\left[\alpha l(\gamma)^{\frac{\eta-1}{\eta}} + (1-\alpha)m(\gamma)^{\frac{\eta-1}{\eta}}\right]^{\frac{1}{\eta-1}}(1-\alpha)m(\gamma)^{-\frac{1}{\eta}}\frac{\sigma}{\sigma-1}m(\gamma)^{\frac{1}{\sigma}}m(\gamma,\zeta)^{\frac{\sigma-1}{\sigma}} - P(\gamma,\zeta)$$

As a point of remark, the infinitesimal firm i takes as given the aggregate quantities that appear in the term:

$$A_{\gamma}\gamma^{\theta}\theta y(l(\gamma),m(\gamma))^{\theta-1}\left[\alpha l(\gamma)^{\frac{\eta-1}{\eta}} + (1-\alpha)m(\gamma)^{\frac{\eta-1}{\eta}}\right]^{\frac{1}{\eta-1}}(1-\alpha)m(\gamma)^{-\frac{1}{\eta}}m(\gamma)^{\frac{1}{\sigma}}.$$

To simplify notation, define:

$$\tau(\gamma) = A_{\gamma}\gamma^{\theta}\theta y(l(\gamma),m(\gamma))^{\theta-1}\left[\alpha l(\gamma)^{\frac{\eta-1}{\eta}} + (1-\alpha)m(\gamma)^{\frac{\eta-1}{\eta}}\right]^{\frac{1}{\eta-1}}(1-\alpha)m(\gamma)^{-\frac{1}{\eta}}m^{-\frac{1}{\eta}}\frac{\sigma}{\sigma-1}m(\gamma)^{\frac{1}{\sigma}}.$$

Here, $\tau(\gamma)$ represents the buyer's heterogeneous valuation for the good sold by the seller. Using the guess that prices scale linearly with the quantity bought, $\tau(\gamma)$ is a strictly increasing function of respective productivities $\gamma = z, \zeta'$. In particular, under the guess, notice that the term:

$$A_{\gamma}\gamma^{\theta}\theta y(l(\gamma),m(\gamma))^{\theta-1}\left[\alpha l(\gamma)^{\frac{\eta-1}{\eta}} + (1-\alpha)m(\gamma)^{\frac{\eta-1}{\eta}}\right]^{\frac{1}{\eta-1}}(1-\alpha)m(\gamma)^{-\frac{1}{\eta}} = p_m$$

Since the term on the right of the equality is the marginal revenue product of the bundle of materials, thus equal to its price p_m, and independent of firm heterogeneity γ. Moreover, recall that

$$m(\gamma)^{\frac{1}{\sigma}} = \left(\frac{\gamma}{\gamma_0}\right)^{\frac{\sigma-1}{\sigma}} m(\gamma_0)^{\frac{1}{\sigma}}$$

Thus:

$$\tau(\gamma) = \gamma^{\frac{\sigma-1}{\sigma}}p_m m(\gamma_0)^{\frac{1}{\sigma}}$$

Therefore, $\tau(\gamma)$ is distributed according to a pareto distribution with tail parameter $\rho := \frac{\kappa\sigma}{\sigma-1}$. Notice that $\rho > \sigma$, since $\kappa > \sigma - 1$. Without loss of generality, we can rewrite the

buyer's surplus from transacting with the infinitesimal seller ζ as:

$$\tilde{\Pi}(\tau(\gamma), \zeta) = \tau(\gamma) m(\tau(\gamma), \zeta)^{\frac{\sigma-1}{\sigma}} - P(\tau(\gamma), \zeta).$$

Lemma 5. *Incentive Compatibility*
The Seller Mechanism Design is Incentive Compatible if and only if:

1.

$$\tilde{\Pi}(\tau(\gamma), \zeta) = \tilde{\Pi}(\tau(\gamma_0), \zeta) + \int_{\tau(\gamma_0)}^{\tau(\gamma)} m(\omega, \zeta)^{\frac{\sigma-1}{\sigma}} d\omega$$

2. $m(\tau(\gamma), \zeta)$ *is increasing in* $\tau(\gamma)$

Proof. The same proof as the classical textbook proof applies here, with infinitely many, infinitesimally small suppliers.

The proof proceeds as follows:

- $\Rightarrow$ Assume Incentive Compatibility:

    1. Denote by $\tilde{\Pi}(\tau(\gamma), \tilde{\tau}(\gamma), \zeta)$ the surplus from a buyer of type $\tau(\gamma)$ faking to be $\tilde{\tau}(\gamma)$. By Incentive Compatibility:

    $$\tilde{\Pi}(\tau(\gamma), \zeta) = \max_{\tilde{\tau}(\gamma)} \tilde{\Pi}(\tau(\gamma), \tilde{\tau}(\gamma), \zeta).$$

    By the envelope theorem:

    $$\frac{d\tilde{\Pi}(\tau(\gamma), \zeta)}{d\tau(\gamma)} = m(\tau(\gamma), \zeta)^{\frac{\sigma-1}{\sigma}}.$$

    Thus, integrating:

    $$\tilde{\Pi}(\tau(\gamma), \zeta) = \tilde{\Pi}(\tau(\gamma_0), \zeta) + \int_{\tau(\gamma_0)}^{\tau(\gamma)} m(\omega, \zeta)^{\frac{\sigma-1}{\sigma}} d\omega.$$

    2. Monotonicity: Pick $\tau_H = \tau(\gamma_H)$ and $\tau_L = \tau(\gamma_L)$ with $\tau_H > \tau_L$. By Incentive Compatibility:

    $$\tau_H m(\tau_H, \zeta)^{\frac{\sigma-1}{\sigma}} - P(\tau_H, \zeta) > \tau_H m(\tau_L, \zeta)^{\frac{\sigma-1}{\sigma}} - P(\tau_L, \zeta).$$

This implies:

$$\tau_H\Big(m(\tau_H,\zeta)^{\frac{\sigma-1}{\sigma}} - m(\tau_L,\zeta)^{\frac{\sigma-1}{\sigma}}\Big) > P(\tau_H,\zeta) - P(\tau_L,\zeta).$$

Again, by Incentive Compatibility:

$$P(\tau_H,\zeta) - P(\tau_L,\zeta) > \tau_L\Big(m(\tau_H,\zeta)^{\frac{\sigma-1}{\sigma}} - m(\tau_L,\zeta)^{\frac{\sigma-1}{\sigma}}\Big).$$

Combining these two results:

$$\tau_H\Big(m(\tau_H,\zeta)^{\frac{\sigma-1}{\sigma}} - m(\tau_L,\zeta)^{\frac{\sigma-1}{\sigma}}\Big) > \tau_L\Big(m(\tau_H,\zeta)^{\frac{\sigma-1}{\sigma}} - m(\tau_L,\zeta)^{\frac{\sigma-1}{\sigma}}\Big).$$

This holds if and only if:

$$m(\tau_H,\zeta) > m(\tau_L,\zeta).$$

- $\Leftarrow$ Take $\tau_H = \tau(\gamma_H)$ and $\tau_L = \tau(\gamma_L)$ with $\tau_H > \tau_L$:

$$\begin{aligned}
\tilde{\Pi}(\tau_H,\zeta) &= \tilde{\Pi}(\tau_L,\zeta) + \int_{\tau_L}^{\tau_H} m(\omega,\zeta)^{\frac{\sigma-1}{\sigma}}\,d\omega \\
&> \tilde{\Pi}(\tau_L,\zeta) + \int_{\tau_L}^{\tau_H} m(\tau_L,\zeta)^{\frac{\sigma-1}{\sigma}}\,d\omega \\
&= \tilde{\Pi}(\tau_L,\zeta) + (\tau_H - \tau_L)m(\tau_L,\zeta)^{\frac{\sigma-1}{\sigma}} \\
&= \tilde{\Pi}(\tau_H,\tau_L,\zeta).
\end{aligned}$$

Here, the first line comes from the assumption of the profit function, the second line from monotonicity, and the remainder from rearranging terms. Hence, τ_H doesn't copy τ_L. An analogous proof applies for τ_L not copying τ_H.

□

Note that individual rationality is satisfied by setting $\tilde{\Pi}(\tau(\gamma_0),\zeta) \geq 0$.

The problem of the seller is then:

$$\begin{aligned}
\max_{\substack{\{P(\tau(z),\zeta),m(\tau(z),\zeta)\},\\ \{P(\tau(\zeta'),\zeta),m(\tau(\zeta'),\zeta)\}}} &\mathbb{E}_z\left[P(\tau(z),\zeta) - c(\zeta)\,m(\tau(z),\zeta)\right]M_z \\
&+ \mathbb{E}_{\zeta'}\left[P(\tau(\zeta'),\zeta) - c(\zeta)\,m(\tau(\zeta'),\zeta)\right]M_\zeta
\end{aligned}$$

Subject to the constraints:

1. Incentive compatibility:

$$\tilde{\Pi}(\tau(\gamma),\zeta) = \tilde{\Pi}(\tau(\gamma_0),\zeta) + \int_{\tau(\gamma_0)}^{\tau(\gamma)} m(\omega,\zeta)^{\frac{\sigma-1}{\sigma}} d\omega$$

2. Monotonicity:

$$m(\tau(\gamma_H),\zeta) > m(\tau(\gamma_L),\zeta), \quad \text{for } \tau(\gamma_H) > \tau(\gamma_L).$$

We solve for the relaxed problem (without monotonicity) and then check that it holds ex-post.

We can substitute for the transfer:

$$P(\tau(\gamma),\zeta) = \tau(\gamma)m(\tau(\gamma),\zeta)^{\frac{\sigma-1}{\sigma}} - \tilde{\Pi}(\tau(\gamma),\zeta) = \tau(\gamma)m(\tau(\gamma),\zeta)^{\frac{\sigma-1}{\sigma}} - \tilde{\Pi}(\tau(\gamma_0),\zeta) - \int_{\tau(\gamma_0)}^{\tau(\gamma)} m(\omega,\zeta)^{\frac{\sigma-1}{\sigma}} d\omega.$$

Now the problem becomes:

$$\max_{\substack{\{m(\tau(z),\zeta)\},\\ \{m(\tau(\zeta'),\zeta)\}}} \int_{z_0} \left[\tau(z)m(\tau(z),\zeta)^{\frac{\sigma-1}{\sigma}} - c(\zeta)\, m(\tau(z),\zeta) - \int_{\tau(z_0)}^{\tau(z)} m(\omega,\zeta)^{\frac{\sigma-1}{\sigma}} d\omega - \tilde{\Pi}(\tau(z_0),\zeta)\right] \mu(z) dz M_z$$
$$+ \int_{\zeta_0} \left[\tau(\zeta')m(\tau(\zeta'),\zeta)^{\frac{\sigma-1}{\sigma}} - c(\zeta)\, m(\tau(\zeta'),\zeta) - \int_{\tau(\zeta_0)}^{\tau(\zeta')} m(\omega,\zeta)^{\frac{\sigma-1}{\sigma}} d\omega - \tilde{\Pi}(\tau(\zeta_0),\zeta)\right] \mu(\zeta') d\zeta' M_\zeta.$$

Now we can apply a substitution of variables to express everything in terms of τ. Specifically, recall that the density $\mu(\tau(\gamma))$ follows a Pareto distribution with tail parameter ρ.

Using the Liebnitz rule and integration by parts, the problem becomes:

$$\max_{\substack{\{m(\tau(z),\zeta)\},\\ \{m(\tau(\zeta'),\zeta)\},\\ \tilde{\Pi}(\tau(z_0),\zeta),\\ \tilde{\Pi}(\tau(\zeta_0),\zeta)}} \int_{\tau(z_0)} \left[(\tau(z) - g^{-1}(\tau(z)))m(\tau(z),\zeta)^{\frac{\sigma-1}{\sigma}} - c(\zeta)\, m(\tau(z),\zeta) - \tilde{\Pi}(\tau(z_0),\zeta)\right] d\tau(z) M_z$$
$$+ \int_{\tau(\zeta_0)} \left[(\tau(\zeta') - g^{-1}(\tau(\zeta')))m(\tau(\zeta'),\zeta)^{\frac{\sigma-1}{\sigma}} - c(\zeta)\, m(\tau(\zeta'),\zeta) - \tilde{\Pi}(\tau(\zeta_0),\zeta)\right] d\tau(\zeta') M_\zeta,$$

where $g(\tau)$ is the hazard function of the τ distribution, given by:

$$g(\tau) = \frac{g(\tau)}{1 - G(\tau)}$$

The seller-firm trivially sets $\tilde{\Pi}(\tau(\zeta_0), \zeta)$ and $\tilde{\Pi}(\tau(z_0), \zeta)$ to zero. Moreover, the virtual utility is given by:

$$\tau(\gamma)\left(\frac{\rho - 1}{\rho}\right) > 0 \quad \forall \quad \tau(\gamma), \quad \gamma = z, \zeta.$$

Where the inequality follows from the fact that $\rho > 1$. Since the virtual utility is always positive, even the lowest type contributes positively. As a result, including all types does not tighten the incentive constraints sufficiently to offset this contribution. In other words, there is no benefit in excluding any type. Consequently, the seller-firm finds it optimal to serve the entire distribution of types, from the lowest to the highest.

Intuitively, excluding the lowest type would allow the firm to gain the information rent from the higher types, while losing the rent the seller is getting from the lowest type. The measure $\frac{1-G(\tau_0)}{g(\tau_0)}$ represents the mass of types with value above τ_0 relative to the density at τ_0. Under a Pareto distribution, the firm would forgo a surplus of τ_0 while gaining only $\frac{\tau_0}{\rho} < \tau_0$, meaning the loss exceeds the gain.

Thus, the seller-firm optimally serves all types $\tau(\gamma)$, starting from the lowest type $\tau(\gamma_0)$.

We are now ready to derive the first-order condition with respect to $m(\tau(\gamma), \zeta)$:

$$\tau(\gamma) m(\tau(\gamma), \zeta)^{-\frac{1}{\sigma}} \frac{\sigma - 1}{\sigma} = c(\zeta) \frac{\rho}{\rho - 1}. \tag{9}$$

Substituting the definition of τ:

$$A_z z^\theta \theta y(l(z), m(z))^{\theta-1} \left[\alpha l(z)^{\frac{\eta-1}{\eta}} + (1-\alpha) m(z)^{\frac{\eta-1}{\eta}}\right]^{\frac{1}{\eta-1}} (1-\alpha) m(z)^{-\frac{1}{\eta}} \left(\frac{m(z)}{m(z,\zeta)}\right)^{\frac{1}{\sigma}} = c(\zeta) \frac{\rho}{\rho - 1}. \tag{10}$$

Similarly:

$$A_\zeta \zeta'^\theta \theta y(l(\zeta'), m(\zeta'))^{\theta-1} \left[\alpha l(\zeta')^{\frac{\eta-1}{\eta}} + (1-\alpha) m(\zeta')^{\frac{\eta-1}{\eta}}\right]^{\frac{1}{\eta-1}} m(\zeta')^{-\frac{1}{\eta}} \left(\frac{m(\zeta')}{m(\zeta', \zeta)}\right)^{\frac{1}{\sigma}} = c(\zeta) \frac{\rho}{\rho - 1}. \tag{11}$$

We then confirm the guess that the transfer scales linearly with quantity:

$$P(\gamma,\zeta)' = p(\zeta) = c(\zeta)\frac{\frac{\kappa}{\theta}}{\frac{\kappa}{\theta}-1}. \tag{12}$$

We are left to find the flat fee that makes $\tilde{\Pi}(\tau(z_0),\zeta)$ and $\tilde{\Pi}(\tau(\zeta_0),\zeta)$ equal to zero.

The flat fee is determined by:

$$\begin{aligned}
F_\gamma(\zeta) &= \tau(\gamma_0)m(\tau(\gamma_0),\zeta)^{\frac{\sigma-1}{\sigma}} - p(\zeta)m(\tau(\gamma_0),\zeta) \\
&= \tau(\gamma_0)m(\tau(\gamma_0),\zeta)^{\frac{\sigma-1}{\sigma}} - \frac{\sigma-1}{\sigma}\tau(\gamma_0)m(\tau(\gamma_0),\zeta)^{\frac{\sigma-1}{\sigma}} \\
&= \tau(\gamma_0)m(\tau(\gamma_0),\zeta)^{\frac{\sigma-1}{\sigma}}\left(\frac{1}{\sigma}\right).
\end{aligned}$$

Now, substituting for $\tau(z_0)$ and $\tau(\zeta_0)$:

$$\begin{aligned}
F_z(\zeta) &= A_z z_0^\theta y(l_0,m_0)^{\theta-1}\left[\alpha l_0^{\frac{\eta-1}{\eta}} + (1-\alpha)m_0^{\frac{\eta-1}{\eta}}\right]^{\frac{1}{\eta-1}}(1-\alpha)m_0^{-\frac{1}{\eta}}m(z_0)^{\frac{1}{\sigma}}m(z_0,\zeta)^{\frac{\sigma-1}{\sigma}}\left(\frac{1}{\sigma}\right), \\
F_\zeta(\zeta) &= A_\zeta \zeta_0^\theta y(l_0,m_0)^{\theta-1}\left[\alpha l_0^{\frac{\eta-1}{\eta}} + (1-\alpha)m_0^{\frac{\eta-1}{\eta}}\right]^{\frac{1}{\eta-1}}(1-\alpha)m_0^{-\frac{1}{\eta}}m(\zeta_0)^{\frac{1}{\sigma}}m(\zeta_0,\zeta)^{\frac{\sigma-1}{\sigma}}\left(\frac{1}{\sigma}\right).
\end{aligned}$$

Using the CES properties:

$$m(\gamma,\zeta)^{\frac{\sigma-1}{\sigma}} = m(\gamma)^{\frac{\sigma-1}{\sigma}}\left(\frac{p(\zeta)}{p_m}\right)^{1-\sigma} = m(\gamma)^{\frac{\sigma-1}{\sigma}}\left(\frac{Q(\zeta)}{D_m}\right)^{\frac{\sigma-1}{\sigma}},$$

where $Q(\zeta)$ is the total production of firm ζ and D_m is the total demand for intermediates.

Substituting into the flat fee:

$$\begin{aligned}
F_\gamma(\zeta) &= A_\gamma \gamma_0^\theta y(l_0,m_0)^{\theta-1}\left[\alpha l_0^{\frac{\eta-1}{\eta}} + (1-\alpha)m_0^{\frac{\eta-1}{\eta}}\right]^{\frac{1}{\eta-1}}(1-\alpha)m_0^{-\frac{1}{\eta}}m(\gamma_0)^{\frac{1}{\sigma}}m(\gamma_0)^{\frac{\sigma-1}{\sigma}}\left(\frac{Q(\zeta)}{D_m}\right)^{\frac{\sigma-1}{\sigma}}\left(\frac{1}{\sigma}\right) \\
&= A_\gamma \gamma_0^\theta y(l_0,m_0)^{\theta-1}\left[\alpha l_0^{\frac{\eta-1}{\eta}} + (1-\alpha)m_0^{\frac{\eta-1}{\eta}}\right]^{\frac{1}{\eta-1}}(1-\alpha)m_0^{-\frac{1}{\eta}}m_0\left(\frac{Q(\zeta)}{D_m}\right)^{\frac{\sigma-1}{\sigma}}\left(\frac{1}{\sigma}\right) \\
&= A_\gamma \gamma_0^\theta y(l_0,m_0)^{\theta}\frac{1}{\left(\frac{\alpha}{1-\alpha}\right)\left(\frac{l}{m}\right)^{\frac{\eta-1}{\eta}}+1}\left(\frac{Q(\zeta)}{D_m}\right)^{\frac{\sigma-1}{\sigma}}\left(\frac{1}{\sigma}\right)
\end{aligned}$$

$$= A_\gamma \gamma_0^\theta y(l_0, m_0)^\theta \frac{1}{\left(\frac{\alpha}{1-\alpha}\right)^\eta \left(\frac{p_m}{w}\right)^{\eta-1} + 1} \left(\frac{Q(\zeta)}{D_m}\right)^{\frac{\sigma-1}{\sigma}} \left(\frac{1}{\sigma}\right),$$

where the last line uses the firm cost minimization condition $\frac{l}{m} = \left(\frac{1-\alpha}{\alpha}\frac{w}{p_m}\right)^{-\eta}$.

Denote by F_z and F_ζ the flat fee charged by the *average* supplier $\tilde{\zeta}$. Then:

$$F_\gamma = A_\gamma \gamma_0^\theta y(l_0, m_0)^\theta \frac{1}{\left(\frac{\alpha}{1-\alpha}\right)^\eta \left(\frac{p_m}{w}\right)^{\eta-1} + 1} \left(\frac{1}{\sigma}\right)\left(\frac{1}{M_\zeta}\right). \quad \gamma = z, \zeta \tag{13}$$

Now, we are ready to verify Guess 2 for the revenue function. For the supplier of type ζ, the revenue can be decomposed as follows:

- **Revenue from the linear part of the two-part tariff:**

$$p(\zeta)Q(\zeta) = Q(\zeta)^{\frac{\sigma-1}{\sigma}} p_m D_m^{\frac{1}{\sigma}}.$$

- **Revenue from the flat fee charged to type z buyers:**

$$F_z(\zeta)M_z = F_z Q(\zeta)^{\frac{\sigma-1}{\sigma}} D_m^{\frac{1-\sigma}{\sigma}} M_z.$$

- **Revenue from the flat fee charged to type ζ buyers:**

$$F_h(\zeta)M_\zeta = F_h Q(\zeta)^{\frac{\sigma-1}{\sigma}} D_m^{\frac{1-\sigma}{\sigma}} M_\zeta.$$

Summing the three components of the revenue, we obtain the revenue function of the *upstream* firm:

$$R_\zeta = Q(\zeta)^{\frac{\sigma-1}{\sigma}} \left[p_m D_m^{\frac{1}{\sigma}} + D_m^{\frac{1-\sigma}{\sigma}} F_h M_\zeta + D_m^{\frac{1-\sigma}{\sigma}} F_z M_z \right].$$

Substituting for the production function after cost minimization:

$$R_\zeta = \zeta^{\frac{\sigma-1}{\sigma}} y\Big(l(\zeta), m(\zeta)\Big)^{\frac{\sigma-1}{\sigma}} \left[p_m D_m^{\frac{1}{\sigma}} + D_m^{\frac{1-\sigma}{\sigma}} F_h M_\zeta + D_m^{\frac{1-\sigma}{\sigma}} F_z M_z \right]. \tag{14}$$

Thus, we verify the guess with $\theta = \frac{\sigma-1}{\sigma}$.
Denote $s_y := \frac{p(\zeta_0)y(\zeta_0)}{D_m}$ as the revenue share of upstream firms with the smallest upstream

type ζ_0. Substituting the expression for F_h and solving for A_ζ, we obtain:

$$A_\zeta = \frac{D_m^{\frac{1-\sigma}{\sigma}}\left(p_m D_m + F_z M_z\right)}{1 - s_y \frac{1}{\left(\frac{\alpha}{1-\alpha}\right)^\eta \left(\frac{p_m}{w}\right)^{\eta-1}+1}\left(\frac{1}{\sigma}\right)} > 0.$$

That completes the verification of the guess. The inequality follows from recognizing that:

$$\frac{p_m m}{cy} = \frac{p_m m}{p_m m\left(1-\alpha\right)^{-\eta}\left(\alpha^\eta\left(\frac{W}{p_m}\right)^{1-\eta} + \left(1-\alpha\right)^\eta\right)} = \frac{1}{\left(1-\alpha\right)^{-\eta}\left(\alpha^\eta\left(\frac{W}{p_m}\right)^{1-\eta} + \left(1-\alpha\right)^\eta\right)} = \frac{1}{\left(\frac{\alpha}{1-\alpha}\right)^\eta\left(\frac{p_m}{w}\right)^{\eta-1}+1}$$

represents the firm's cost share of materials.

On the other hand, the revenue function of the *downstream* firm is:

$$R_z = Q(z)^{\frac{\sigma-1}{\sigma}} Y^{\frac{1}{\sigma}} = z^{\frac{\sigma-1}{\sigma}} y\Big(l(z), m(z)\Big)^{\frac{\sigma-1}{\sigma}} Y^{\frac{1}{\sigma}}. \tag{15}$$

Thus, the guesses are verified with the following:

- $\theta = \frac{\sigma-1}{\sigma}$,
- $A_z = Y^{\frac{1}{\sigma}}$,
- $A_\zeta = \frac{D_m^{\frac{1-\sigma}{\sigma}}\left(p_m D_m + F_h M_\zeta\right)}{1-s_y \frac{1}{\left(\frac{\alpha}{1-\alpha}\right)^\eta\left(\frac{p_m}{w}\right)^{\eta-1}+1}\left(\frac{1}{\sigma}\right)}$,
- $p(\zeta) = \frac{\rho}{\rho-1}c(\zeta)$, where $\rho = \frac{\kappa\sigma}{\sigma-1}$.

Proposition 3. *The solution to the mechanism design problem of the seller is described by:*

- $P(\gamma, \zeta) = p(\zeta)m(\gamma, \zeta) + F_\gamma, \quad \gamma = z, \zeta,$
- $p(\zeta) = \frac{\rho}{\rho-1}c(\zeta), \quad \rho = \frac{\sigma\kappa}{\sigma-1},$
- $F_\gamma(\zeta) = Revenue(\gamma_0)\frac{1}{\left(\frac{\alpha}{1-\alpha}\right)^\eta\left(\frac{p_m}{w}\right)^{\eta-1}+1}\left(\frac{1}{\sigma}\right).$

Remarks:

1. The markup on the linear part of the two-part tariff is lower than the linear price markup:

$$\rho > \sigma \Leftrightarrow \kappa > \sigma - 1$$

which is a condition required for finite output. Recall that:

$$TFP = \tilde{z} = \left[\mathbb{E}_z z^{\sigma-1}\right]^{\frac{1}{\sigma-1}}.$$

2. Profits of the lowest type **are not** zero. To see this, consider the downstream firm z_0. It is well known that in the context of CES demand, variable profits are a fraction $\frac{1}{\sigma}$ of revenue, while variable costs take a fraction $\frac{\sigma-1}{\sigma}$ of revenues. Moreover, the fraction of the linear part of intermediate costs out of total cost is:

$$\frac{p_m m}{cy} = \frac{p_m m}{p_m m\,(1-\alpha)^{-\eta}\left(\alpha^{\eta}\left(\frac{W}{p_m}\right)^{1-\eta} + (1-\alpha)^{\eta}\right)} = \frac{1}{(1-\alpha)^{-\eta}\left(\alpha^{\eta}\left(\frac{W}{p_m}\right)^{1-\eta} + (1-\alpha)^{\eta}\right)}.$$

Note that when $\eta \to 1$ this collapse to the usual Cobb-Douglas share $(1-\alpha)$.

Thus:

$$p_m m = \frac{\sigma-1}{\sigma}\text{Revenue}\frac{1}{(1-\alpha)^{-\eta}\left(\alpha^{\eta}\left(\frac{W}{p_m}\right)^{1-\eta} + (1-\alpha)^{\eta}\right)}.$$

Summing the revenue parts that go to variable profits, the linear price of the two-part tariff, the flat fee, and labor costs:

$$\begin{aligned}\Pi = \text{Revenue}\Bigg[1 &- \frac{1}{(1-\alpha)^{-\eta}\left(\alpha^{\eta}\left(\frac{W}{p_m}\right)^{1-\eta} + (1-\alpha)^{\eta}\right)}\left(\frac{1}{\sigma}\right) + \\ &- \frac{\sigma-1}{\sigma}\frac{1}{(1-\alpha)^{-\eta}\left(\alpha^{\eta}\left(\frac{W}{p_m}\right)^{1-\eta} + (1-\alpha)^{\eta}\right)} + \\ &- \frac{\sigma-1}{\sigma}\left(1 - \frac{1}{(1-\alpha)^{-\eta}\left(\alpha^{\eta}\left(\frac{W}{p_m}\right)^{1-\eta} + (1-\alpha)^{\eta}\right)}\right)\Bigg].\end{aligned}$$

Denote s_m as the intermediate share of total variable costs. The profit of the lowest

type z_0 is given by:

$$\Pi(z_0) = \text{Revenue}(z_0)\left[1 - \frac{1}{\sigma}s_m - \frac{\sigma-1}{\sigma}\right]. \tag{16}$$

In particular, it's easy to verify that when $s_m = 0$, profits revert to a share $\frac{1}{\sigma}$ of total revenue, while when the intermediate share s_m approaches one, total profit approaches zero. This indicates that the lowest-type buyer receives some rent due to their ability to substitute intermediates with labor.

3. **Average Profits:** Denote by $s_m := \frac{1}{(1-\alpha)^{-\eta}\left(\alpha^\eta\left(\frac{W}{p_m}\right)^{1-\eta}+(1-\alpha)^\eta\right)}$ and by $w = \frac{W}{p_m}$. Note, that:

$$F_z M_\zeta = \frac{p_m s_m}{\sigma-1}\left((1-\alpha)^\eta + \alpha^\eta w^{1-\eta}\right)^{\frac{1}{1-\eta}} \Upsilon \frac{M_z^{\frac{\sigma}{\sigma-1}}}{\tilde{z}}\left(\frac{z_0}{\tilde{z}}\right)^{\sigma-1}$$

The average profits for the downstream firms is given by:

$$\begin{aligned}\mathbb{E}\left[\Pi(z)\right] &= p_m \frac{1}{\sigma-1}\left((1-\alpha)^\eta + \alpha^\eta w^{1-\eta}\right)^{\frac{1}{1-\eta}} \Upsilon \frac{M_z^{\frac{\sigma}{1-\sigma}}}{\tilde{z}} - M_z \\ &= p_m \frac{1}{\sigma-1}\left((1-\alpha)^\eta + \alpha^\eta w^{1-\eta}\right)^{\frac{1}{1-\eta}} \Upsilon \frac{M_z^{\frac{\sigma}{1-\sigma}}}{\tilde{z}}\left(\frac{\tilde{z}^{\sigma-1} - s_m z_0^{\sigma-1}}{\tilde{z}^{\sigma-1}}\right)\end{aligned} \tag{17}$$

The average profits for the upstream firm is given by:

$$\begin{aligned}\mathbb{E}\left[\Pi(\zeta)\right] &= p_m \frac{1}{\rho-1}\left((1-\alpha)^\eta + \alpha^\eta w^{1-\eta}\right)^{\frac{1}{1-\eta}} D_m \frac{M_\zeta^{\frac{\sigma}{1-\sigma}}}{\tilde{\zeta}} + F_z M_z + F_h M_\zeta - M_{z=} \\ &= p_m \frac{1}{\rho-1}\left((1-\alpha)^\eta + \alpha^\eta w^{1-\eta}\right)^{\frac{1}{1-\eta}} D_m \frac{M_\zeta^{\frac{\sigma}{1-\sigma}}}{\tilde{\zeta}} + \frac{M_z}{M_\zeta} p_m s_m \frac{1}{\sigma-1}\left((1-\alpha)^\eta + \alpha^\eta w^{1-\eta}\right)^{\frac{1}{1-\eta}} \Upsilon \frac{M_z^{\frac{\sigma}{1-\sigma}}}{\tilde{z}} \frac{z_0^{\sigma-1}}{\tilde{z}^\sigma}\end{aligned} \tag{18}$$

If you also assume $\tilde{\zeta} = \tilde{z}$, then:

$$\mathbb{E}\left[\Pi(\zeta)\right] = p_m\left((1-\alpha)^\eta + \alpha^\eta w^{1-\eta}\right)^{\frac{1}{1-\eta}} \frac{1}{\tilde{z}}\left(\frac{D_m M_\zeta^{\frac{\sigma}{1-\sigma}}}{\rho-1} + \frac{s_m \Upsilon M_z^{\frac{1}{1-\sigma}}}{\sigma-1}\frac{z_0^{\sigma-1}}{\tilde{z}^{\sigma-1}}\right)$$

B.4 Checking that All Types Are Served

Consider z downstream firms, (for the types ζ' is specular). For ease of notation, we express everything in terms of the buyer's valuation $\tau(z)$ rather than z. One way of checking that all types are served is to substitute for the solution for assuming the lowest type is served, and then evaluating monotonicity of the resulting profit function for increases in the lowest type.

The seller's profit function is given by:

$$\Pi = \int_{\tau_0} \left[(p(\zeta) - c(\zeta)) m(\tau, \zeta) + F \right] g(\tau)\, d\tau.$$

We seek to verify that the seller does not have an incentive to exclude the lowest type τ_0. This requires showing that:

$$\frac{d\Pi}{d\tau_0} < 0.$$

Recall the expressions for the linear pricing component and the fixed fee:

$$p(\zeta) = \frac{\rho}{\rho - 1} c(\zeta),$$

$$F(\zeta) = \tau_0 m(\tau_0, \zeta)^{\frac{\sigma-1}{\sigma}} \frac{1}{\sigma},$$

where $\rho = \frac{\sigma\kappa}{\sigma-1}$, with σ representing the elasticity of substitution across varieties and κ denoting the tail index of the Pareto distribution. Moreover, recall the quantity allocated at the optimal solution:

$$m(\tau, \zeta) = \left[c(\zeta) \frac{\rho}{\rho - 1} \frac{\sigma}{\sigma - 1} \frac{1}{\tau} \right]^{-\sigma}.$$

Thus, total profits are given by:

$$\Pi = \int_{\tau_0} \left[\frac{1}{\rho - 1} c \left[c \frac{\rho}{\rho - 1} \frac{\sigma}{\sigma - 1} \frac{1}{\tau} \right]^{-\sigma} + \tau_0 \left[c \frac{\rho}{\rho - 1} \frac{\sigma}{\sigma - 1} \frac{1}{\tau_0} \right]^{1-\sigma} \frac{1}{\sigma} \right] g(\tau)\, d\tau.$$

Using the Leibniz integral rule, we obtain:

$$\frac{d\Pi}{d\tau_0} = -\frac{1}{\rho - 1} c \left[c \frac{\rho}{\rho - 1} \frac{\sigma}{\sigma - 1} \frac{1}{\tau_0} \right]^{-\sigma} g(\tau_0) + F' - F(\tau_0) g(\tau_0).$$

Recall that $g(\tau)$ is a Pareto with tail parameter $\frac{\kappa}{\theta}$. Then, expanding further:

$$\frac{d\Pi}{d\tau_0} = -\frac{1}{\rho-1}c\left[c\frac{\rho}{\rho-1}\frac{\sigma}{\sigma-1}\frac{1}{\tau_0}\right]^{-\sigma}\frac{\kappa\theta^{-1}}{\tau_0}$$

$$+\sigma\tau_0^{\sigma-1}\left[c\frac{\rho}{\rho-1}\frac{\sigma}{\sigma-1}\right]^{1-\sigma}\frac{1}{\sigma} - \tau_0^{\sigma-1}\left[c\frac{\rho}{\rho-1}\frac{\sigma}{\sigma-1}\right]^{1-\sigma}\frac{1}{\sigma}(\kappa\theta^{-1}).$$

Rewriting, we obtain:

$$\frac{d\Pi}{d\tau_0} = -\frac{1}{\rho-1}\tau_0^{\sigma-1}c\left[c\frac{\rho}{\rho-1}\frac{\sigma}{\sigma-1}\right]^{-\sigma}\frac{\kappa}{\theta}$$

$$+\tau_0^{\sigma-1}\left[c\frac{\rho}{\rho-1}\frac{\sigma}{\sigma-1}\right]^{1-\sigma}\left[1-\frac{\kappa\theta^{-1}}{\sigma}\right].$$

Since $\theta = \frac{\sigma-1}{\sigma}$, it follows that:

$$1-\frac{\kappa\theta^{-1}}{\sigma} = 1-\frac{\kappa}{\sigma-1} < 0,$$

where the last inequality holds because $\kappa > \sigma - 1$. This implies that:

$$\frac{d\Pi}{d\tau_0} < 0,$$

confirming that the seller optimally serves all types, including the lowest type τ_0.

C Model details

C.1 Environment

We will simplify the model towards proving existence and uniqueness, but all proofs remain valid for the original model. We assume that $\eta = 1$ from our main CES production function specification so that firm production functions are Cobb-Douglas.

$$y(\gamma) = l^{\alpha} m^{1-\alpha}, \quad \gamma \in \{z, \zeta\}$$

Where m is a CES aggregator of upstream firms varieties:

$$m(\gamma) = \left(\int_{\zeta_0} m(\gamma, \zeta)^{\frac{\sigma-1}{\sigma}} \, M_{\zeta} \, \mu(\zeta) \, d\zeta \right)^{\frac{\sigma}{\sigma-1}}, \quad \gamma \in \{z, \zeta\}$$

p_m represents the CES price index of the intermediate bundle.

$$p_m = \left(\int_{\zeta_0} p_{\zeta}^{1-\sigma} M_{\zeta} \mu_{\zeta} d\zeta \right)^{\frac{1}{1-\sigma}}$$

The corresponding marginal cost function is:

$$c(\gamma; w, p_m) = \frac{1}{\gamma} w^{\alpha} p_m^{1-\alpha} \Theta, \quad \Theta = \left(\alpha^{\alpha} (1-\alpha)^{1-\alpha} \right)^{-1}, \quad \gamma \in \{z, \zeta\}$$

Moreover:

$$\frac{l}{m} = \left(\frac{\alpha}{1-\alpha} \frac{p_m}{W} \right)$$

The production function can also be rewritten as:

$$y_{\gamma} = \gamma l \left(\frac{1-\alpha}{\alpha} \frac{w}{p_m} \right)^{1-\alpha}, \quad \gamma \in \{z, \zeta\}$$

C.2 Efficient Pricing Equilibrium

Firms must be allowed to charge a markup sufficient to incentivize the optimal entry level. But markup distorts input choices by effectively acting as a uniform tax on production. To mitigate this distortion, an output subsidy is required to restore marginal-cost pricing, conditional on entry.

In particular, firms receive a subsidy on production τ so that their perceived marginal cost is equal to $\tilde{c} = \frac{\sigma-1}{\sigma}c$ (i.e. $\tau = \frac{\sigma-1}{\sigma}$). The subsidy is paid via a lump sum tax to the household.

Downstream firms profits

Then, the equilibrium conditions, as in the linear pricing benchmark, are detailed as follows. Starting with the profit function for a downstream firm:

$$\begin{aligned}
\pi_z &= (p_z - \tilde{c})y_z \\
&= \left(\frac{\sigma}{\sigma-1}\tilde{c} - \tilde{c}\right)y_z \\
&= \left(\frac{1}{\sigma-1}\right)\tilde{c}y_z \\
&= \left(\frac{1}{\sigma-1}\right)\tau c y_z \\
&= \left(\frac{1}{\sigma-1}\right)\tau\frac{1}{z}w^{\alpha}p_m^{1-\alpha}\Theta y_z
\end{aligned}$$

From CES demand:

$$y_z = Y\left(\frac{p_z}{P}\right)^{-\sigma}$$

The price index with symmetric firms is:

$$\begin{aligned}
P &= \left(M_z\int p_z^{1-\sigma}\mu(z)dz\right)^{\frac{1}{1-\sigma}} \\
&= M_z^{\frac{1}{1-\sigma}}p_z
\end{aligned}$$

Substituting back:

$$\begin{aligned}
\pi_z &= \left(\frac{1}{\sigma-1}\right)\tau\frac{1}{z}w^{\alpha}p_m^{1-\alpha}\Theta Y\left(\frac{p_z}{M_z^{\frac{1}{1-\sigma}}p_z}\right)^{-\sigma} \\
&= \left(\frac{1}{\sigma-1}\right)\tau\frac{1}{z}w^{\alpha}p_m^{1-\alpha}\Theta Y M_z^{\frac{\sigma}{1-\sigma}}
\end{aligned}$$

Taking expectations:

$$\begin{aligned} E[\pi_z] &= \frac{1}{\sigma-1}\tau w^{\alpha} p_m^{1-\alpha}\Theta Y M_z^{\frac{\sigma}{1-\sigma}} E\left[\frac{1}{z}\right] \\ &= \frac{1}{\sigma-1}\tau w^{\alpha} p_m^{1-\alpha}\Theta Y \frac{M_z^{\frac{\sigma}{1-\sigma}}}{\tilde{z}} \end{aligned}$$

Hence, the zero profit condition for downstream firms is:

$$\frac{1}{\sigma-1}\tau w^{\alpha} p_m^{1-\alpha}\Theta Y \frac{M_z^{\frac{\sigma}{1-\sigma}}}{\tilde{z}} = (1-\delta)c_e w.$$

Zero profit condition for upstream firms: Denote D_ζ as the aggregate demand for intermediates. Then:

$$\frac{1}{\sigma-1}\tau w^{\alpha} p_m^{1-\alpha}\Theta D_\zeta \frac{M_\zeta^{\frac{\sigma}{1-\sigma}}}{\tilde{\zeta}} = (1-\delta)c_e w.$$

Production equals demand for the average productivity $\tilde{\zeta}$ type of upstream firm. Denote l_ζ the labor used by average upstream firm $\tilde{\zeta}$ (i.e. $\mathbb{E}l(z)$):

$$\tilde{\zeta} l_\zeta \left(\frac{1-\alpha}{\alpha}\frac{W}{p_m}\right)^{1-\alpha} = \left(\frac{p(\tilde{z})}{P}\right)^{-\sigma} D_\zeta = D_\zeta M_\zeta^{\frac{\sigma}{1-\sigma}}.$$

Same condition for the downstream firms

$$\tilde{z} l_z \left(\frac{1-\alpha}{\alpha}\frac{W}{p_m}\right)^{1-\alpha} = Y M_z^{\frac{\sigma}{1-\sigma}}.$$

Labor market clearing. Denote $M = M_z + M_\zeta$:

$$1 = (1-\delta)Mc_e + M_\zeta l_\zeta + M_z l_z.$$

The optimal pricing with a subsidy for upstream firms is:

$$\begin{aligned} p_\zeta &= \frac{\sigma}{\sigma-1}\tilde{c}_\zeta \\ &= \frac{\sigma}{\sigma-1}c_\zeta \tau \\ &= \frac{\sigma}{\sigma-1}\tau\frac{1}{\zeta}w^{\alpha} p_m^{1-\alpha}\Theta \end{aligned}$$

Substituting this into the price index:

$$p_m = \left(\int_{\zeta_0} \left(\frac{\sigma}{\sigma - 1} \tau \frac{1}{\zeta} w^\alpha p_m^{1-\alpha} \Theta \right)^{1-\sigma} M_\zeta \mu_\zeta d\zeta \right)^{\frac{1}{1-\sigma}}$$
$$= \frac{\sigma}{\sigma - 1} \tau w^\alpha p_m^{1-\alpha} \Theta \left(M_\zeta \int_{\zeta_0} \zeta^{\sigma-1} \mu_\zeta d\zeta \right)^{\frac{1}{1-\sigma}}$$

Using the definition of average productivity: $\tilde{\zeta} = \left(\int_{\zeta_0} \zeta^{\sigma-1} \mu_z dz \right)^{\frac{1}{\sigma-1}}$:

$$p_m = \frac{\sigma}{\sigma - 1} M_\zeta^{\frac{1}{1-\sigma}} \frac{1}{\tilde{\zeta}} w^\alpha p_m^{1-\alpha} \Theta \tau$$

Which implies:

$$M_\zeta^{\frac{1}{1-\sigma}} = \tilde{\zeta} \frac{\sigma - 1}{\sigma} \frac{p_m^\alpha}{w^\alpha \Theta \tau}$$

Applying the same logic for downstream firms:

$$M_z^{\frac{1}{1-\sigma}} = \tilde{z} \frac{\sigma - 1}{\sigma} \frac{p_m^{\alpha-1}}{w^\alpha \Theta \tau}$$

Lastly, as labor can go to the production of both upstream and downstream firms or to finance entry, denote $L = l_z M_z + l_\zeta M_\zeta = 1 - (1 - \delta) c_e M$. Then, aggregate aggregate demand of intermediates:

$$D_\zeta = \frac{w}{p_m} \left(\frac{1 - \alpha}{\alpha} \right) (l_z M_z + l_\zeta M_\zeta) = \frac{w}{p_m} \left(\frac{1 - \alpha}{\alpha} \right) L.$$

Solving the System

We have a system of 8 equations and 8 unknowns $M_z, M_\zeta, p_m, w, l_z, l_\zeta, D_\zeta, Y$

zero profit condition downstream:

$$\frac{1}{\sigma - 1} \tau w^\alpha p_m^{1-\alpha} \Theta Y \frac{M_z^{\frac{\sigma}{1-\sigma}}}{\tilde{z}} = (1 - \delta) c_e w \tag{19}$$

zero profit condition upstream:

$$\frac{1}{\sigma-1}\tau w^{\alpha} p_m^{1-\alpha}\Theta D_{\zeta}\frac{M_{\zeta}^{\frac{\sigma}{1-\sigma}}}{\tilde{\zeta}} = (1-\delta)c_e w \tag{20}$$

Production equals demand for the average productivity type of downstream firm:

$$\tilde{z}l_z\left(\frac{1-\alpha}{\alpha}\frac{w}{p_m}\right)^{1-\alpha} = YM_z^{\frac{\sigma}{1-\sigma}} \tag{21}$$

Production equals demand for the average productivity type of upstream firm:

$$\tilde{\zeta}l_{\zeta}\left(\frac{1-\alpha}{\alpha}\frac{w}{p_m}\right)^{1-\alpha} = D_{\zeta}M_{\zeta}^{\frac{\sigma}{1-\sigma}} \tag{22}$$

Labor market clearing:

$$1 = (1-\delta)Mc_e + M_{\zeta}l_{\zeta} + M_z l_z \tag{23}$$

Price index downstream:

$$M_z^{\frac{1}{1-\sigma}} = \tilde{z}\frac{\sigma-1}{\sigma}\frac{p_m^{\alpha-1}}{w^{\alpha}\Theta\tau} \tag{24}$$

Price index upstream:

$$M_{\zeta}^{\frac{1}{1-\sigma}} = \tilde{\zeta}\frac{\sigma-1}{\sigma}\frac{p_m^{\alpha}}{w^{\alpha}\Theta\tau} \tag{25}$$

Intermediate demand equals supply:

$$D_{\zeta} = \frac{w}{p_m}\left(\frac{1-\alpha}{\alpha}\right)L \tag{26}$$

Towards proving existence and uniqueness, we will assume that mean productivities for upstream and downstream firms are equivalent, $\tilde{\zeta} = \tilde{z}$. Taking the ratio of equations 19 and 20:

$$\begin{aligned} Y\frac{M_z^{\frac{\sigma}{1-\sigma}}}{\tilde{z}} &= D_{\zeta}\frac{M_{\zeta}^{\frac{\sigma}{1-\sigma}}}{\tilde{\zeta}} \\ YM_z^{\frac{\sigma}{1-\sigma}} &= D_{\zeta}M_{\zeta}^{\frac{\sigma}{1-\sigma}} \end{aligned}$$

Using this and combining it with equations 21 and 22:

$$l_z = l_{\zeta} = l$$

Now plug in equation 21 into 19 and using the fact that $\tau = \frac{\sigma-1}{\sigma}$:

$$\begin{aligned}
\frac{1}{\sigma-1}\tau w^{\alpha} p_m^{1-\alpha}\Theta Y \frac{M_z^{\frac{\sigma}{1-\sigma}}}{\tilde{z}} &= (1-\delta)c_e w \\
\frac{1}{\sigma-1}\tau w^{\alpha} p_m^{1-\alpha}\Theta \frac{1}{\tilde{z}} \tilde{z} l_z \left(\frac{1-\alpha}{\alpha}\frac{w}{p_m}\right)^{1-\alpha} &= (1-\delta)c_e w \\
\frac{1}{\sigma-1}\tau w \Theta l \left(\frac{1-\alpha}{\alpha}\right)^{1-\alpha} &= (1-\delta)c_e w \\
\frac{1}{\sigma-1}\tau l \alpha^{-1} &= (1-\delta)c_e \\
l &= (\sigma-1)(1-\delta)c_e \alpha \tau^{-1} \\
l &= \sigma(1-\delta)c_e \alpha
\end{aligned}$$

Now, take equation and plug it in equation 20:

$$\begin{aligned}
\frac{1}{\sigma-1}\tau w^{\alpha} p_m^{1-\alpha}\Theta D_{\zeta} \frac{M_{\zeta}^{\frac{\sigma}{1-\sigma}}}{\tilde{\zeta}} &= (1-\delta)c_e w \\
\frac{1}{\sigma-1}\tau w^{\alpha} p_m^{1-\alpha}\Theta \frac{w}{p_m}\left(\frac{1-\alpha}{\alpha}\right) L \frac{M_{\zeta}^{\frac{\sigma}{1-\sigma}}}{\tilde{\zeta}} &= (1-\delta)c_e w \\
\frac{1}{\sigma-1}\tau w^{\alpha+1} p_m^{-\alpha}\Theta \left(\frac{1-\alpha}{\alpha}\right) L \frac{M_{\zeta}^{\frac{\sigma}{1-\sigma}}}{\tilde{\zeta}} &= (1-\delta)c_e w
\end{aligned}$$

Use this equation and take the ratio of equation 25 to the above equation and solving for $M\zeta$:

$$\begin{aligned}
M_{\zeta} &= \tilde{\zeta}\frac{\sigma-1}{\sigma}\frac{p_m^{\alpha}}{w^{\alpha}\Theta\tau}\frac{1}{(1-\delta)c_e w}\frac{1}{\sigma-1}\tau w^{\alpha+1} p_m^{-\alpha}\Theta\left(\frac{1-\alpha}{\alpha}\right)L\frac{1}{\tilde{\zeta}} \\
M_{\zeta} &= \left(\frac{1-\alpha}{\alpha}\right)\frac{1}{\sigma}\frac{L}{(1-\delta)c_e}
\end{aligned}$$

Substituting this equation back into equation 25, using the fact that $\tau = \frac{\sigma-1}{\sigma}$ and solving for L

$$\begin{aligned}
M_{\zeta}^{\frac{1}{1-\sigma}} &= \tilde{\zeta}\frac{\sigma-1}{\sigma}\frac{p_m^{\alpha}}{w^{\alpha}\Theta\tau} \\
M_{\zeta} &= \left[\tilde{\zeta}\frac{\sigma-1}{\sigma}\frac{p_m^{\alpha}}{w^{\alpha}\Theta\tau}\right]^{1-\sigma}
\end{aligned}$$

$$\left(\frac{1-\alpha}{\alpha}\right)\frac{1}{\sigma}\frac{L}{(1-\delta)c_e} = \left[\tilde{\zeta}\frac{\sigma-1}{\sigma}\frac{p_m^{\alpha}}{w^{\alpha}\Theta\tau}\right]^{1-\sigma}$$

$$L = \underbrace{\left[\tilde{\zeta}\frac{1}{\Theta}\right]^{1-\sigma}\left(\frac{\alpha}{1-\alpha}\right)\sigma(1-\delta)c_e}_{F}\underbrace{\left(\frac{w}{p_m}\right)^{\alpha(\sigma-1)}}_{\Phi(\Omega)}$$

$$L = F\Phi(\Omega)$$

Where Ω is the ratio wage over CES intermediate inputs price ratio. Using the fact that labor goes to production or to pay for entry:

$$l(\Omega) = 1 - (1-\delta)Mc_e$$
$$M(\Omega) = \frac{1 - F\Phi(\Omega)}{(1-\delta)c_e}$$

We reduced the system to two equations:

$$l(\Omega) = \sigma(1-\delta)c_e\alpha \tag{27}$$
$$M(\Omega) = \frac{1 - F\Phi(\Omega)}{(1-\delta)c_e} \tag{28}$$

To find the solution, we must obtain Ω, using the labor market clearing condition:

$$1 = (1-\delta)c_eM + Ml(\Omega)$$
$$1 = M\left[(1-\delta)c_e + l(\Omega)\right]$$
$$1 = \frac{1 - F\Phi(w)}{(1-\delta)c_e}\left[(1-\delta)c_e + l(\Omega)\right]$$
$$\frac{1}{(1-\delta)c_e + l(\Omega)} = \frac{1 - F\Phi(w)}{(1-\delta)c_e}$$

Define

$$H(\Omega) := \frac{1}{(1-\delta)c_e + l(\Omega)} - \frac{1}{(1-\delta)c_e} + \frac{F\Phi(\Omega)}{(1-\delta)c_e}$$
$$= g(\Omega) - \frac{1}{(1-\delta)c_e} + \frac{F\Phi(\Omega)}{(1-\delta)c_e}$$

The equilibrium Ω^* is such that $H(\Omega^*) = 0$.

Proposition 1 (Unique Equilibrium). There is a unique $\Omega^* > 0$ such that $g(\Omega^*) =$

$M(\Omega^*)$.

Proof. **Step 1: Continuity** $H(\Omega)$ is a continuous function. It comes straightforward from being a sum of continuous functions.

Step 2: Limits As $\Omega \to 0$, $\Phi(\Omega) \to 0$. Thus

$$H(\Omega) = \frac{1}{(1-\delta)c_e + (1-\delta)c_e(\sigma)\alpha} - \frac{1}{(1-\delta)c_e} < 0$$

Moreover:

$$\lim_{\Omega \to \infty} M(\Omega) = -\infty$$

Since $\lim_{\Omega\to\infty} \Phi(\Omega) = \infty$. Thus

$$\lim_{\Omega \to \infty} H(\Omega) = \infty$$

Step 3: Existence Since $H(0) < 0$, and $H(\infty) > 0$ and since $H(\Omega)$ is continuous it must intersect 0 at least once by intermediate value theorem.

Step 4: Uniqueness $\Phi(\Omega)$ is either a strictly convex function or a strictly concave depending on whether $(\sigma - 1) > \frac{1}{\alpha}$. Thus $H(\Omega)$ is either a strictly convex or concave function. Thus the solution is unique.

□

Once Ω is known, M_z is obtain through :

$$M_\zeta(\Omega) = \left(\tilde{z}\frac{\sigma-1}{\sigma}\right)^{1-\sigma} \Omega^{(\sigma-1)\alpha}\Theta^{\sigma-1}\tau^{\sigma-1}$$

$$M_\zeta^{\frac{1}{1-\sigma}} D_\zeta = \tilde{z}\frac{\sigma-1}{\sigma}\Theta^{-1}\tau^{-1}\Omega^{-\alpha}\left(\frac{1-\alpha}{\alpha}\right)\Omega L$$

Using 22, notice that:

$$\tilde{z} l_\zeta M_\zeta \left(\frac{1-\alpha}{\alpha}\Omega\right)^{1-\alpha} = D_\zeta M_\zeta^{\frac{1}{1-\sigma}} = \tilde{z}\frac{\sigma-1}{\sigma}\Theta^{-1}\tau^{-1}\Omega^{-\alpha}\left(\frac{1-\alpha}{\alpha}\right)\Omega L$$

Simplifying:

$$l_\zeta M_\zeta = s_m L$$

$$L = (l_z M_z + l_\zeta M_\zeta)$$

Which implies

$$l_\zeta M_\zeta = \frac{s_m}{1 - s_m} l_z M_z$$

C.3 Linear Price Equilibrium

Following the efficient pricing problem without a subsidy, the system of equations is characterized as follows.

zero profit condition downstream:

$$\frac{1}{\sigma - 1} w^\alpha p_m^{1-\alpha} \Theta Y \frac{M_z^{\frac{\sigma}{1-\sigma}}}{\tilde{z}} = (1 - \delta) c_e w \tag{29}$$

zero profit condition upstream:

$$\frac{1}{\sigma - 1} w^\alpha p_m^{1-\alpha} \Theta D_\zeta \frac{M_\zeta^{\frac{\sigma}{1-\sigma}}}{\tilde{\zeta}} = (1 - \delta) c_e w \tag{30}$$

Production equals demand for the average productivity type of downstream firm:

$$\tilde{z} l_z \left(\frac{1 - \alpha}{\alpha} \frac{w}{p_m} \right)^{1-\alpha} = Y M_z^{\frac{\sigma}{1-\sigma}} \tag{31}$$

Production equals demand for the average productivity type of upstream firm:

$$\tilde{\zeta} l_\zeta \left(\frac{1 - \alpha}{\alpha} \frac{w}{p_m} \right)^{1-\alpha} = D_\zeta M_\zeta^{\frac{\sigma}{1-\sigma}} \tag{32}$$

Labor market clearing:

$$1 = (1 - \delta) M c_e + M_\zeta l_\zeta + M_z l_z \tag{33}$$

Price index downstream:

$$M_z^{\frac{1}{1-\sigma}} = \tilde{z} \frac{\sigma - 1}{\sigma} \frac{p_m^{\alpha-1}}{w^\alpha \Theta} \tag{34}$$

Price index upstream:

$$M_\zeta^{\frac{1}{1-\sigma}} = \tilde{\zeta} \frac{\sigma - 1}{\sigma} \frac{p_m^{\alpha}}{w^\alpha \Theta} \tag{35}$$

Intermediate demand equals supply:

$$D_\zeta = \frac{w}{p_m}\left(\frac{1-\alpha}{\alpha}\right)L \tag{36}$$

Using a homologous strategy than in the efficient pricing equilibrium, we reduce the system to arrive at the following two equations:

$$l(\Omega) = \sigma(1-\delta)c_e(\sigma-1)\alpha \tag{37}$$

$$L(\Omega) = \left(\tilde{z}\frac{\sigma-1}{\sigma}\right)^{1-\sigma}\sigma(1-\delta)c_e\left(\frac{1-\alpha}{\alpha}\right)^{-1}\Omega^{(\sigma-1)\alpha}\Theta^{(\sigma-1)}$$
$$L(\Omega) = FK_{LP}\Phi(\Omega)$$

Where F is as before:

$$F = (\tilde{z})^{1-\sigma}\,\sigma(1-\delta)c_e\,(1-\alpha)^{-1}\,\alpha\Theta^{(\sigma-1)}$$

And K_{LP} is a new constant equal to:

$$K_{LP} = \left(\frac{\sigma-1}{\sigma}\right)^{1-\sigma} > 1$$

And

$$\Phi(\Omega) = \Omega^{(\sigma-1)\alpha}$$

Using $1-(1-\delta)M = L$:

$$M(\Omega) = \frac{1-FK_{LP}\Phi(\Omega)}{(1-\delta)c_e}$$

To find the solution, we must obtain Ω, using the labor market clearing condition:

$$1 = (1-\delta)c_eM + Ml = M\left((1-\delta)c_e + l\right)$$

Define

$$H(\Omega) := \frac{1}{(1-\delta)c_e + l(\Omega)} - \frac{1}{(1-\delta)c_e} + \frac{FK_{LP}\Phi(\Omega)}{(1-\delta)c_e}$$
$$= g(\Omega) - \frac{1}{(1-\delta)c_e} + \frac{FK_{LP}\Phi(\Omega)}{(1-\delta)c_e}$$

The equilibrium Ω^* is such that $H(\Omega^*) = 0$.

Proposition 1 (Unique Equilibrium). There is a unique $\Omega^* > 0$ such that $g(\Omega^*) = M(\Omega^*)$.

Proof. **Step 1: Continuity** $H(\Omega)$ is a continuous function. It comes straightforward from being a sum of continuous functions.

Step 2: Limits As $\Omega \to 0$, $\Phi(\Omega) \to 0$. Thus

$$H(\Omega) = \frac{1}{(1-\delta)c_e + (1-\delta)c_e(\sigma-1)\alpha} - \frac{1}{(1-\delta)c_e} < 0$$

Moreover:

$$\lim_{\Omega\to\infty} M(\Omega) = -\infty$$

Since $\lim_{\Omega\to\infty} \Phi(\Omega) = \infty$. Thus

$$\lim_{\Omega\to\infty} H(\Omega) = \infty$$

Step 3: Existence Since $H(0) < 0$, and $H(\infty) > 0$ and since $H(\Omega)$ is continuous it must intersect 0 at least once by intermediate value theorem.

Step 4: Uniqueness $\Phi(\Omega)$ is either a strictly convex function or a strictly concave depending on whether $(\sigma - 1) > \frac{1}{\alpha}$. Thus $H(\Omega)$ is either a strictly convex or concave function. Thus the solution is unique.

□

C.4 Nonlinear Price Equilibrium

Let's recall the equilibrium conditions. Let $s_m := 1 - \alpha$ be the intermediate expenditure share of total costs. We will now use the notation $\Omega = \frac{w}{p}$ from the initial equations. The equilibrium conditions are summarized by:

Zero profit condition for downstream firms:

$$\frac{1}{\sigma-1}\Omega^\alpha \Theta Y \frac{M_z^{\frac{\sigma}{1-\sigma}}}{\tilde{z}} - \frac{1}{p_m} F_z M_\zeta = (1-\delta)c_e\Omega$$

Which becomes:

$$\frac{1}{\sigma-1}\Omega^\alpha \Theta Y \frac{M_z^{\frac{\sigma}{1-\sigma}}}{\tilde{z}} \left(\frac{\tilde{z}^{\sigma-1} - s_m z_0^{\sigma-1}}{\tilde{z}^{\sigma-1}}\right) = (1-\delta)c_e\Omega. \tag{38}$$

Zero profit condition for upstream firms: Denote D_ζ as the aggregate demand for intermediates. Then:

$$\frac{1}{\rho-1}\Omega^\alpha\Theta D_\zeta\frac{M_\zeta^{\frac{\sigma}{1-\sigma}}}{\tilde{z}}+\frac{1}{p_m}(F_zM_z-F_\zeta M_\zeta+F_\zeta M_\zeta)=(1-\delta)c_e\Omega$$

Which becomes:

$$\Omega^\alpha\Theta\frac{1}{\tilde{z}}\left[\frac{D_\zeta M_\zeta^{\frac{\sigma}{1-\sigma}}}{\rho-1}+\frac{M_z}{M_\zeta}\frac{z_0^{\sigma-1}}{\tilde{z}^{\sigma-1}}\frac{s_mYM_z^{\frac{\sigma}{1-\sigma}}}{\sigma-1}\right]=(1-\delta)c_e\Omega. \tag{39}$$

Production equal demand for the average productivity $\tilde{z}$ type of upstream firm. Denote l_ζ the labor used by average upstream firm $\tilde{z}$ (i.e. $\mathbb{E}l(z)$):

$$\tilde{z}l_\zeta\left(\frac{1-\alpha}{\alpha}\Omega\right)^{1-\alpha}=D_\zeta M_\zeta^{\frac{\sigma}{1-\sigma}}. \tag{40}$$

Same condition for the downstream firms

$$\tilde{z}l_z\left(\frac{1-\alpha}{\alpha}\Omega\right)^{1-\alpha}=YM_z^{\frac{\sigma}{1-\sigma}}. \tag{41}$$

Labor market clearing. Denote $M=M_z+M_\zeta$:

$$1=(1-\delta)Mc_e+M_\zeta l_\zeta+M_zl_z. \tag{42}$$

Using the definition of the CES price index of intermediates we can arrive to:

$$M_\zeta^{\frac{1}{1-\sigma}}=\tilde{z}\frac{\rho-1}{\rho}\Omega^{-\alpha}\Theta^{-1} \tag{43}$$

Lastly, aggregate demand of intermediates:

$$D_\zeta=\Omega\left(\frac{1-\alpha}{\alpha}\right)L. \tag{44}$$

C.5 Solving The System

Denote $s = (\frac{z_0}{\tilde{z}})^{\sigma-1}$. From 43 we already have a function $M_\zeta(\Omega)$:

$$M_\zeta(\Omega) = \left(\tilde{z}\frac{\rho-1}{\rho}\right)^{1-\sigma} \Omega^{(\sigma-1)\alpha}\Theta^{\sigma-1}$$

Denote by L total labor used in production (i.e. $L = M_z l_z + M_\zeta l_\zeta = 1 - (1-\delta)Mc_e$). Using 44 and 43:

$$M_\zeta^{\frac{1}{1-\sigma}} D_\zeta = \tilde{z}\frac{\rho-1}{\rho}\Theta^{-1}\Omega^{-\alpha}\left(\frac{1-\alpha}{\alpha}\right)\Omega L$$

Using 40, notice that:

$$\tilde{z} l_\zeta M_\zeta \left(\frac{1-\alpha}{\alpha}\Omega\right)^{1-\alpha} = D_\zeta M_\zeta^{\frac{1}{1-\sigma}} = \tilde{z}\frac{\rho-1}{\rho}\Theta^{-1}\Omega^{-\alpha}\left(\frac{1-\alpha}{\alpha}\right)\Omega L$$

Simplifying:

$$l_\zeta M_\zeta = \frac{\rho-1}{\rho} s_m L$$

Denote $q = \frac{\rho-1}{\rho} s_m$

$$L = (l_z M_z + l_\zeta M_\zeta)$$

Which implies

$$l_\zeta M_\zeta = \frac{q}{1-q} l_z M_z$$

Also, taking the ratio of (40) and (41), we know

$$l_\zeta = \frac{D_\zeta M_\zeta^{\frac{\sigma}{1-\sigma}}}{Y M_z^{\frac{\sigma}{1-\sigma}}} l_z$$

Combining the two:

$$\frac{M_z}{M_\zeta} = \frac{1-q}{q}\frac{D_\zeta M_\zeta^{\frac{\sigma}{1-\sigma}}}{Y M_z^{\frac{\sigma}{1-\sigma}}}$$

Thus we can rewrite the zero profit condition (39):

$$\Omega^\alpha \Theta \frac{1}{\tilde{z}} D_\zeta M_\zeta^{\frac{\sigma}{1-\sigma}} \left[\frac{1}{\rho-1} + \frac{1-q}{q}\frac{s s_m}{\sigma-1}\right] = (1-\delta)c_e\Omega.$$

Substituting for the value of $q(\Omega)$ and simplifying:

$$\Omega^{\alpha}\Theta\frac{1}{\tilde{z}}D_{\zeta}M_{\zeta}^{\frac{\sigma}{1-\sigma}}\frac{1}{(\rho-1)(\sigma-1)}\left[\sigma-1+\rho s-(\rho-1)s_m s\right]=(1-\delta)c_e\Omega.$$

Substitute for D_{ζ} in this equation:

$$\Omega^{\alpha}\Theta\frac{1}{\tilde{z}}\Omega\left(\frac{1-\alpha}{\alpha}\right)LM_{\zeta}^{\frac{\sigma}{1-\sigma}}\frac{1}{(\rho-1)(\sigma-1)}\left[\sigma-1+\rho s-(\rho-1)s_m(\Omega)s\right]=(1-\delta)c_e\Omega.$$

$$M_{\zeta}^{\frac{\sigma}{1-\sigma}}=(1-\delta)c_e(\rho-1)(\sigma-1)\tilde{z}\left(1-\alpha\right)^{-1}\alpha\frac{\Theta^{-1}\Omega^{-\alpha}}{\left[\sigma-1+\rho s-(\rho-1)s_m s\right]}\frac{1}{L}$$

Also recall:

$$M_{\zeta}^{\frac{1}{1-\sigma}}=\tilde{z}\frac{\rho-1}{\rho}\Omega^{-\alpha}\Theta^{-1}$$

Taking the ratio of the second equation with the first:

$$M_{\zeta}=\left(\frac{1-\alpha}{\alpha}\right)\frac{1}{\rho}\frac{1}{(1-\delta)c_e(\sigma-1)}L\left[\sigma-1+\rho s-(\rho-1)s_m s\right]$$

Using equation (43) we can solve for $L(\Omega)$:

$$\left(\tilde{z}\frac{\rho-1}{\rho}\right)^{1-\sigma}\Omega^{(\sigma-1)\alpha}\Theta^{\sigma-1}=\left(\frac{1-\alpha}{\alpha}\right)\frac{1}{\rho}\frac{1}{(1-\delta)c_e(\sigma-1)}L\left[\sigma-1+\rho s-(\rho-1)s_m s\right]$$

From which:

$$L(\Omega)=FK_{NLP}\Phi(\Omega)$$

Where F is as before

$$F=(\tilde{z})^{1-\sigma}\left(\frac{\alpha}{1-\alpha}\right)\sigma c_e(1-\delta)\Theta^{\sigma-1}$$

And

$$K_{NLP}=\left(\frac{\rho-1}{\rho}\right)^{1-\sigma}\frac{\rho}{\sigma}\frac{\sigma-1}{\left[\sigma-1+\rho s-(\rho-1)s_m s\right]}>1$$

And

$$\Phi(\Omega)=\Omega^{(\sigma-1)\alpha}$$

Knowing L we can solve for $M(\Omega)$:

$$1 = (1-\delta)c_e M + L \rightarrow M(\Omega) = \frac{1 - FK_{NLP}\Phi(\Omega)}{(1-\delta)c_e}$$

Now recall zero profit condition for upstream firms:

$$\Omega^{\alpha}\Theta\frac{1}{\tilde{z}}D_{\zeta}M_{\zeta}^{\frac{\sigma}{1-\sigma}}\frac{1}{(\rho-1)(\sigma-1)}\left[\sigma - 1 + \rho s - (\rho-1)s_m s\right] = (1-\delta)c_e\Omega.$$

Recall zero profit condition for downstream:

$$\frac{1}{\sigma-1}\Omega^{\alpha}\Theta Y\frac{M_z^{\frac{\sigma}{1-\sigma}}}{\tilde{z}}\left(\frac{\tilde{z}^{\sigma-1} - s_m z_0^{\sigma-1}}{\tilde{z}^{\sigma-1}}\right) = (1-\delta)c_e\Omega.$$

Moreover recall:

$$\tilde{z}l_{\zeta}\left(\frac{\Omega(1-\alpha)}{\alpha}\right)^{1-\alpha} = D_{\zeta}M_{\zeta}^{\frac{\sigma}{1-\sigma}}.$$

Same condition for the downstream firms

$$\tilde{z}l_{z}\left(\frac{\Omega(1-\alpha)}{\alpha}\right)^{1-\alpha} = YM_{z}^{\frac{\sigma}{1-\sigma}}.$$

Thus we can rewrite the zero profit conditions as:

$$\Omega^{\alpha}\Theta\frac{1}{\tilde{z}}\tilde{z}l_{\zeta}\left(\frac{\Omega(1-\alpha)}{\alpha}\right)^{1-\alpha}\frac{1}{(\rho-1)(\sigma-1)}\left[\sigma - 1 + \rho s - (\rho-1)s_m s\right] = (1-\delta)c_e\Omega.$$

and

$$\frac{1}{\sigma-1}\Omega^{\alpha}\Theta l_{z}\left(\frac{\Omega(1-\alpha)}{\alpha}\right)^{1-\alpha}(1 - s_m s) = (1-\delta)c_e\Omega.$$

Taking the ratio:

$$\frac{l_{\zeta}}{l_z}\frac{1}{\rho-1}\frac{\sigma - 1 + \rho s - (\rho-1)s_m s}{1 - s s_m} = 1$$

From which we can solve for $l(\Omega)$:

$$l_{\zeta}(\Omega) = l_z(\Omega)(\rho-1)(1-ss_m)\frac{1}{\sigma - 1 + \rho s - (\rho-1)s_m s}$$

And

$$l_z = (1-\delta)c_e\alpha(\sigma-1)\frac{1}{1-ss_m}$$

Also

$$\frac{D_\zeta M_\zeta^{\frac{\sigma}{1-\sigma}}}{YM_z^{\frac{\sigma}{1-\sigma}}} = \frac{l_\zeta}{l_z}$$

Thus:

$$\frac{M_z}{M_\zeta} = \frac{1-q}{q}\frac{D_\zeta M_\zeta^{\frac{\sigma}{1-\sigma}}}{YM_z^{\frac{\sigma}{1-\sigma}}} = \frac{1-q}{q}\frac{l_\zeta}{l_z}$$

$$M_z(\Omega) = \frac{1-q}{q}\theta M_\zeta(\Omega)$$

Where $\theta(\Omega) = \frac{l_\zeta}{l_z}$:

$$\theta(\Omega) = \frac{(\rho-1)(1-ss_m)}{\sigma-1+\rho s-(\rho-1)ss_m}$$

Notice that:

$$M = M_\zeta + M_z = M_\zeta\left(1+\frac{1-q}{q}\theta\right)$$

And notice that:

$$l_z(\Omega)M_z(\Omega) + l_\zeta(\Omega)M_\zeta(\Omega) = l_z(\Omega)\frac{1-q}{q}\theta M_\zeta(\Omega) + \theta l_z M_\zeta(\Omega) = \theta l_z M_\zeta(\Omega)\left(1+\frac{1-q}{q}\right)$$

Thus the labor market clearing condition can be rewritten as:

$$1 = (1-\delta)c_e M_\zeta\left(1+\frac{1-q)}{q}\theta\right) + \theta l_z M_\zeta\left(1+\frac{1-q}{q}\right)$$

With:

$$M_\zeta(\Omega) = \frac{1-FK_{NLP}\Phi(\Omega)}{(1-\delta)c_e}\left(1+\frac{1-q}{q}\theta\right)^{-1}$$

The equilibrium Ω^* is such that labor market clears. This means that the two function intersect:

$$g = M(\Omega)$$

Where

$$g = \left[(1-\delta)c_e + \theta l_z\left(1+\frac{1-q}{q}\theta\right)^{-1}\left(1+\frac{1-q}{q}\right)\right]^{-1}$$

$$M(\Omega) = \frac{1 - FK_{NLP}\Phi(\Omega)}{(1-\delta)c_e}$$

And

$$M_\zeta(\Omega) = \frac{1 - FK_{NLP}\Phi(\Omega)}{(1-\delta)c_e}\left(1 + \frac{1-q}{q}\theta\right)^{-1}$$

Equilibrium Ω^* is such that the following function crosses zero:

$$H(\Omega) = g - M_\zeta(\Omega)$$

There exists a unique solution

Proof. As $\Omega \to 0$, $\lim_{\Omega\to 0}\Phi(\Omega) = 0$. Thus

$$\lim_{\Omega\to 0} M_\zeta(\Omega) = \left[(1-\delta)c_e\left(1 + \frac{1-q}{q}\theta\right)\right]^{-1} > g$$

Thus, $H(0) < 0$. Moreover, as for the linear case, $\Phi(\Omega)$ is a strictly increasing function that goes to ∞ as $\Omega \to \infty$. Moreover, it is either strictly concave or strictly convex depending on whether $\sigma - 1 > \frac{1}{\alpha}$.

Existence: Since $H(0) < 0$ and $\lim_{\Omega\to\infty} H(\Omega) = \infty$, and since $H(\Omega)$ is a continuous function, $H(\Omega)$ crosses zero at least once.

Uniqueness: Since $M_\zeta(\Omega)$ is either strictly concave or convex, also $H(\Omega)$ is either concave or convex. Thus the solution is unique □