DOCUMENTOS DE TRABAJO

The Aggregate Welfare Effects of Nonlinear Prices in Supply Chains

Luca Lorenzini
Antonio Martner

N° 1049 Julio 2025
BANCO CENTRAL DE CHILE

banco central Chile

Documentos de Trabajo del Banco Central de Chile
Working Papers of the Central Bank of Chile
Agustinas 1180, Santiago, Chile
Teléfono: (56-2) 3882475; Fax: (56-2) 38822311

Documento de Trabajo N° 1049

Working Paper N° 1049

The Aggregate Welfare Effects of Nonlinear Prices in Supply Chains*

Luca Lorenzini
UCLA Anderson

Antonio Martner
UCLA and Central Bank of Chile

Resumen

Este trabajo estudia como el poder de mercado de empresas vendedoras imponiendo precios no lineales en cadenas productivas afecta la cantidad producida, la entrada de empresas y el bienestar agregado. Desarrollamos un modelo de equilibrio general en el que las empresas tanto cobran como pagan precios no lineales a lo largo de la cadena productiva. En comparación con precios lineales, precios no lineales aumentan la cantidad producida a nivel de empresa pero reducen la entrada de firmas al distorsionar la distribución de beneficios entre firmas, lo que genera efectos ambiguos sobre el bienestar agregado. Utilizando datos a nivel de transacción de empresas chilenas, documentamos evidencia consistente con una práctica generalizada de vendedores imponiendo precios no lineales a distintos grupos de compradores. Al calibrar el modelo con los datos, encontramos que los precios no lineales elevan la cantidad producida pero desalientan la entrada de firmas. En equilibrio, predominan las ganancias en producción: las pérdidas de bienestar agregadas asociadas al poder de mercado de vendedores son aproximadamente un 18% menores bajo precios no lineales que bajo precios lineales, lo que indica que los análisis basados en precios lineales sobrestiman los costos de bienestar agregado del poder de mercado.

Abstract

We study how nonlinear pricing in supply chains shapes output, firm entry, and aggregate welfare. We develop a general equilibrium model in which firms both charge and pay nonlinear prices along the supply chain. Relative to linear pricing, nonlinear prices increase firm-level output but reduce firm entry by distorting the distribution of profits, yielding ambiguous welfare effects. Using transaction-level data from Chilean firms, we document robust evidence consistent with widespread nonlinear pricing across buyer groups. Calibrating the model to the data, we find that nonlinear pricing raises production but deters entry. In equilibrium, output gains dominate: aggregate welfare losses from market power are approximately 18% lower under nonlinear prices than under linear pricing, indicating that analyses based on linear pricing overstate 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, Ariel Burstein, Jonathan Vogel, Federico Huneeus, Yasutaka Koike-Mori and Mounu Prem 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

Policymakers have expressed growing concern about rising market power and its consequences for economy-wide efficiency and aggregate welfare. Yet market power alone does not necessarily generate inefficiencies or welfare losses, what matters is how that power is exercised through pricing. Standard models typically assume uniform pricing, where a single price applies to all units sold.

However, decades of research in industrial organization show that firms routinely deviate from this assumption. As Varian (1989) observed nearly four decades ago, "Price discrimination is one of the most prevalent forms of marketing practices," noting that "there can be no doubt that firms are well aware of the benefits of price discrimination." Wilson (1993) similarly highlighted the ubiquity of nonlinear pricing: "What do phone rates, frequent flyer programs, and railroad tariffs all have in common? They are all examples of nonlinear pricing."

Using transaction-level data covering the universe of formal firm-to-firm transactions in Chile, we find widespread evidence consistent with price discrimination. Firm-to-firm trade accounts for approximately 8 times firm-to-consumer trade in Chile's formal economy. Given this scale, the way market power is exercised through pricing in supply chains may have sizable implications for aggregate efficiency and welfare.

In this paper, we develop a supply chain model in which firms both charge and face nonlinear prices when buying and selling inputs. Compared to a linear pricing framework, nonlinear pricing improves allocations but alters the distribution of rents across firms, which in turn distorts firm entry. Using transaction-level firm-to-firm data from Chile, we find evidence inconsistent with uniform pricing and find indicative evidence of nonlinear pricing across buyer groups. Guided by these empirical patterns, we calibrate the model and find that nonlinear pricing raises firm-level production but discourages firm entry relative to linear pricing. In equilibrium, the production gains dominate: aggregate welfare losses from market power are approximately 18% lower under nonlinear prices than under linear pricing. This suggests that models assuming linear pricing may overstate the aggregate welfare costs of market power.

We outline this paper's theoretical framework before turning to the empirical evidence and quantification. We develop a two-step theoretical framework. First, we solve a monopolistic screening problem where a seller faces buyers with productivities following a Pareto distribution. Under homothetic revenue and constant marginal cost, the optimal contract is a two-part tariff: a constant marginal price and a seller-specific flat fee based on

the buyer distribution. Second, we embed this framework into a multitier supply chain. The structure ensures full buyer participation and incentive compatibility, allowing recursive implementation across layers. This scalability makes the model suitable for studying nonlinear pricing in general equilibrium within supply chains.

In nonlinear pricing, observed average unit prices do not map directly to quantities. The flat fee extracts surplus without affecting marginal decisions, while the constant marginal price determines allocations. This allocative price depends on demand elasticity and the Pareto tail parameter, and its markup is strictly lower than the markup under linear pricing. As a result, quantity distortions are smaller. In supply chains, allocative prices act like output taxes, reducing input use and output, but less so than under uniform pricing.

Nonlinear pricing alters the distribution of rents relative to linear pricing, which in turn distorts firm entry. We distinguish retailers, which buy inputs and sell to final consumers, from upstream firms, which both buy and sell inputs and form a full supply chain among themselves. Nonlinear pricing applies only to transactions between upstream firms and from upstream firms to retailers. Retailers pay flat fees to upstream firms, reducing their profits while leaving variable margins unchanged. Upstream firms both pay and collect flat fees, with the most productive firms benefiting most. This asymmetric redistribution compresses returns for marginal entrants, especially among retailers, and distorts entry patterns across the supply chain.

Nonlinear pricing introduces a tradeoff between improved allocations and distorted entry. On the intensive margin, lower allocative markups reduce distortions in input use and output, despite flat fees compressing firm scale. On the extensive margin, flat fees act as implicit entry taxes for retailers, while upstream firms' payments and receipts cancel in expectation. In equilibrium, firm-level output increases but entry declines. Restoring efficiency requires three instruments: output subsidies to eliminate marginal distortions, lump sum transfers to offset rent shifts, and entry subsidies to correct participation incentives.

We use administrative records covering the universe of firm-to-firm transactions in Chile's formal economy, capturing every transaction with product, price, quantity, buyer, and seller identifiers. This level of granularity, across all sectors and firms, provides a rare view into supply chain pricing at scale. A core empirical challenge is distinguishing second from third-degree price discrimination, as prices vary with both quantities and buyer characteristics. Large buyers tend to purchase more and differ systematically

by sector, size, or region, making it difficult to separate nonlinear pricing from targeted pricing. Sellers may also mix posted prices with negotiated terms, further blurring the underlying pricing mechanisms.

We find departures from uniform pricing in 71% of transactions, a reality ignored by most macroeconomic models, and find that quantity and buyer observables together explain 43% of residual price variation, after controlling for daily demand and supply shocks and ensuring product-unit comparability. Prices decline nonlinearly with quantity, consistent with second-degree price discrimination, while systematic variation across buyer groups points to third-degree discrimination. Controlling for buyer fixed effects increases estimated quantity discounts, contradicting the idea that steeper discounts reflect buyer power. A proxy for buyer power—the number of suppliers a buyer transacts with—shows no meaningful interaction with pricing. These results support a seller-driven pricing model, consistent with our framework where firms design nonlinear contracts across buyer groups.

We calibrate the model using Chilean firm size distribution data and parameter values from the literature, then validate it against observed pricing patterns. The calibrated model reproduces the nonlinear pricing patterns in the data without targeting them directly, matching the shape and magnitude of quantity discounts. The model also captures heterogeneity across buyer groups and aligns with the observed relationship between prices, quantities, and firm characteristics.

Nonlinear pricing achieves 66% of efficient welfare, compared to 59% under linear pricing, reducing welfare losses by 18%. This improvement occurs despite added pricing complexity, including asymmetric entry distortions and heterogeneous markups along the supply chain. Both regimes feature excessive entry, as markup-driven rents attract marginal firms beyond the efficient level. Welfare losses are primarily driven by intensive margin distortions, accounting for 63% of the total. Among these, material usage inefficiencies dominate, contributing 65–71% of total losses through compounded markups. While extensive margin distortions explain 37%, misallocation from double marginalization—distorting input use and output—remains the main channel through which market power reduces welfare. Our main result is that assuming linear pricing overstates the welfare costs of market power: under nonlinear pricing, entry is reduced but allocations improve.

Our findings suggest that competition policy should consider the efficiency effects of pricing restrictions when market power is present. While price discrimination raises

traditional concerns about rent extraction and market access, our results indicate that allowing nonlinear pricing can improve welfare by enhancing allocative efficiency, even when it reduces firm entry. Regulatory frameworks that restrict pricing flexibility may face a trade-off between limiting rent extraction and preserving allocation benefits. In settings with substantial market power, policymakers may need to weigh these efficiency gains against distributional concerns about how pricing practices affect heterogeneous firms and market participation.

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 nonlinear prices 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 general equilibrium model in which firms charge and pay nonlinear prices.

Second, we engage with the literature on supply chains by incorporating more 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 A Model of Nonlinear Prices in Supply Chains

We develop a theoretical framework to analyze the welfare effects of nonlinear pricing in supply chains. Building on canonical mechanism design theory, we show that when firm productivities follow Pareto distributions, homothetic revenue functions, and constant marginal costs, profit-maximizing sellers optimally implement nonlinear price contracts based on a two-part tariffs that can be characterized in closed form. These contracts feature quantity discounts through marginal prices below monopolistic competition with CES demand systems levels, combined with flat fees that extract buyer surplus.

When extended to multi-layer supply chains, this pricing structure generates the following welfare tradeoff: lower marginal prices improve allocative efficiency by reducing double marginalization, but flat fees create entry distortions by redistributing surplus across firm types. We characterize the conditions under which nonlinear pricing improves aggregate welfare relative to uniform pricing and identify the policy instruments needed to achieve first-best allocations.

2.1 The Optimal Nonlinear Price: Basic Construct

We analyze the canonical monopolistic screening problem to develop scalable optimal nonlinear pricing in supply chains. Under Pareto-distributed buyer types, homothetic revenue functions, and constant marginal costs, the optimal nonlinear price takes the form of a two-part tariff with seller-specific flat fees and a constant marginal price. The solution exhibits the properties that enable recursive implementation across supply chain layers, with two results in equilibrium: all buyer types participate, and sellers have no incentive to deviate from the optimal schedule for any buyer.

Primitives. We consider a standard mechanism design problem where a monopolist seller interacts with a continuum of buyers indexed by their type z. Buyer types represent productivity levels that determine their valuation for the seller's product, and this productivity is private information not observable by the seller.

Buyer types z are distributed Pareto with shape parameter $\kappa > 1$ and support $[1, \infty)$. The lowest type has $z_{\min} = 1$, so the probability density and cumulative distribution functions are:

$$f(z) = \kappa z^{-\kappa-1}, \quad F(z) = 1 - z^{-\kappa}.$$

For simplicity, assume that the seller has a constant marginal cost c, and that each buyer has the following homothetic revenue function[1], which is increasing in buyer type z:

$$R(z,q) = \frac{z^{\sigma-1}q^{\frac{\sigma-1}{\sigma}}}{(\sigma-1)/\sigma},$$

where σ is the price-demand elasticity of buyers for the seller's good.

Seller Behavior. The seller can choose to price discriminate by offering a menu of contracts, each specifying a transfer $T(z)$ and a quantity $q(z)$ for each buyer type z. Define the total mass of buyers as N_z, so that the seller's mechanism design profit maximization problem is:

$$\begin{aligned}\max_{\{T(z),q(z)\}} \Pi = \mathbb{E}_z\left[T(z) - cq(z)\right]N_z \quad \text{s.t.}\\ \text{(IR)} \quad \Pi(z) \geq 0, \quad \forall z\\ \text{(IC)} \quad R(z,q(z)) - T(z) \geq R(z,q(\tilde{z})) - T(\tilde{z}), \quad \forall z,\tilde{z}\end{aligned}$$

The individual rationality (IR) constraints ensure that each buyer receives nonnegative surplus from transacting with the seller, while the incentive compatibility (IC) constraints ensure that no buyer prefers the contract designed for another buyer type. We assume that arbitrage is not possible and that quantities are monotonic in types, meaning that higher types purchase larger quantities and have uniformly higher willingness to pay.

The Optimal Nonlinear Price. Applying the Revelation Principle, we can restrict attention to direct truthful mechanisms where all buyers correctly reveal their private information. Each buyer's profit is given by:

$$\Pi(z) = R(z,q(z)) - T(z),$$

To ensure that buyers truthfully reveal their type, we impose the incentive compatibility (IC) condition. By the Envelope Theorem, if IC holds and the profit function is

[1]This functional form yields closed-form solutions, but the results extend to any homothetic revenue function.

differentiable, then:

$$\Pi'(z) = \frac{\partial R(z, q(z))}{\partial z} = \frac{(\sigma - 1)z^{\sigma-2}}{\gamma} q(z)^{1-\frac{1}{\sigma}},$$

where $\gamma = (\sigma - 1)/\sigma$.

Integrating and normalizing $\Pi(1) = 0$ for the lowest type, we obtain:

$$\Pi(z) = \int_1^z \frac{(\sigma - 1)\tilde{z}^{\sigma-2}}{\gamma} q(\tilde{z})^{1-1/\sigma} d\tilde{z}.$$

This condition ensures that higher types derive higher profits, which requires that $q(z)$ be non-decreasing in z. From the profit function definition, we can express the transfer $T(z)$ as:

$$T(z) = R(z, q(z)) - \Pi(z) = \frac{z^{\sigma-1}}{\gamma} q(z)^{1-\frac{1}{\sigma}} - \Pi(z).$$

The monopolist's expected profit is:

$$\begin{aligned} \Pi_{\text{Seller}} &= \int_1^\infty \left[T(z) - cq(z)\right] f(z) dz \\ &= \int_1^\infty \left[\frac{z^{\sigma-1}}{\gamma} q(z)^{1-\frac{1}{\sigma}} - \Pi(z) - cq(z)\right] f(z) dz. \end{aligned}$$

Substituting the expression for $\Pi(z)$ into the integral and applying integration by parts to eliminate the information rent term, we derive:

$$\Pi_{\text{Seller}} = \int_1^\infty \left[\frac{z^{\sigma-1}}{\gamma} - \frac{1 - F(z)}{f(z)} \cdot \frac{d}{dz}\left(\frac{z^{\sigma-1}}{\gamma}\right) - cq(z)\right] q(z)^{1-1/\sigma} f(z) dz.$$

The expression inside the brackets is known as the virtual valuation. It adjusts the buyer's marginal valuation downward to account for the information rent that must be left to the buyer to induce truthful revelation. For the Pareto distribution, we can compute the hazard rate:

$$f(z) = \kappa z^{-\kappa-1}, \quad F(z) = 1 - z^{-\kappa}, \quad \Rightarrow \quad h(z) = \frac{f(z)}{1 - F(z)} = \frac{\kappa}{z} \quad \Rightarrow \frac{1 - F(z)}{f(z)} = \frac{z}{\kappa}.$$

We compute the virtual valuation term:

$$\frac{z^{\sigma-1}}{\gamma} - \frac{z}{\kappa} \cdot \frac{d}{dz}\left(\frac{z^{\sigma-1}}{\gamma}\right) = \frac{z^{\sigma-1}}{\gamma} - \frac{z}{\kappa} \cdot \frac{(\sigma-1)z^{\sigma-2}}{\gamma} = \frac{z^{\sigma-1}}{\gamma}\left(1 - \frac{\sigma-1}{\kappa}\right).$$

Define the parameter ρ which will govern pricing as:

$$\rho = \frac{\sigma\kappa}{\sigma-1} \quad \Rightarrow \quad 1 - \frac{\sigma-1}{\kappa} = \frac{\kappa-(\sigma-1)}{\kappa} = \frac{\rho-\sigma}{\rho}.$$

Thus, the virtual valuation becomes:

$$\frac{z^{\sigma-1}}{\gamma} \cdot \frac{\rho-\sigma}{\rho}.$$

The monopolist chooses $q(z)$ to maximize pointwise profit:

$$\max_{q(z)} \left\{ \frac{z^{\sigma-1}}{\gamma} \cdot \frac{\rho-\sigma}{\rho} q(z)^{1-1/\sigma} - cq(z) \right\}.$$

The first-order condition yields:

$$\left(1 - \frac{1}{\sigma}\right) \cdot \frac{z^{\sigma-1}}{\gamma} \cdot \frac{\rho-\sigma}{\rho} q(z)^{-1/\sigma} = c.$$

Solving for $q(z)$ gives:

$$q(z) = \left[\left(1 - \frac{1}{\sigma}\right) \cdot \frac{z^{\sigma-1}}{\gamma c} \cdot \frac{\rho-\sigma}{\rho}\right]^{\sigma}.$$

This quantity is positive when $\kappa > \sigma$, ensuring that all buyer types participate. Higher types receive larger quantities, with the elasticity ρ determining how steeply quantity increases with buyer productivity. This elasticity reflects both the demand elasticity σ and the shape of the productivity distribution κ.

We now show that this optimal mechanism can be implemented using a two-part tariff of the form:

$$T(z) = F + pq(z),$$

where F is a flat fee and p is the per-unit price. The buyer's profit is:

$$\Pi(z) = \frac{z^{\sigma-1}}{\gamma} q(z)^{1-1/\sigma} - pq(z).$$

The buyer chooses $q(z)$ to maximize profit, leading to the first-order condition:

$$p = \frac{\partial R(z,q)}{\partial q} = \frac{z^{\sigma-1}}{\gamma}\left(1 - \frac{1}{\sigma}\right) q(z)^{-1/\sigma}.$$

Substituting the seller-optimal quantity $q(z)$ into the buyer's first-order condition, we find:

$$p = \frac{\rho}{\rho - 1} c.$$

The monopolist sets the per-unit price above marginal cost c, extracting rents from private information about buyer types. The flat fee F extracts buyer surplus while maintaining incentive compatibility and individual rationality. The flat fee equals the surplus of the lowest type, whose individual rationality constraint binds.

Proposition 1. ***Optimal Two-Part Tariff***
Under Pareto-distributed buyer types with shape parameter $\kappa > \sigma$ and homothetic revenue functions, constant marginal costs, the optimal nonlinear pricing mechanism can be implemented as a two-part tariff consisting of:

- ***Per-unit price:***
$$p^* = \frac{\rho}{\rho - 1} c, \quad \textit{where } \rho = \frac{\sigma\kappa}{\sigma - 1}.$$
- ***Flat fee:***
$$F = \frac{z_{\min}^{\sigma-1}}{\gamma} q(z_{\min})^{1-\frac{1}{\sigma}} - p^* q(z_{\min}),$$
where $z_{\min} = 1$ is the lowest buyer type.

This implements the optimal allocation, ensures incentive compatibility, prevents arbitrage, and extracts maximum profit given asymmetric information.

Two results from this optimal pricing structure are worth highlighting. First, it is always optimal for the seller to serve all buyer types. While excluding the lowest type $z_{\min} = 1$ would allow the seller to charge a higher flat fee, doing so would result in the loss

of demand from that type. Under the Pareto distribution, the virtual valuation remains strictly positive for all types, implying that even the lowest type contributes positively to seller profits[2].As a result, including all types does not tighten the incentive constraints sufficiently to justify exclusion. There is no benefit to excluding any type (see Appendix B.1 for more details).

Second, there is no profitable deviation by the seller from the optimal price schedule to any buyer type. An important robustness result from Wilson (1993) is that the seller has no incentive to deviate from the optimal nonlinear price schedule by charging different prices for different buyer types or quantities. Once the two-part tariff is designed optimally, the marginal price $p(q) = T'(q)$ already aligns incentives across the demand curve.

The seller could, in principle, charge a different marginal price p for each unit q, potentially extracting more surplus from certain types. However, Wilson's heuristic demonstrates that such deviations are not profitable. Using the inverse demand function derived from buyer optimization:

$$z(q, p) = q^{1/(\sigma-1)} p^{\sigma/(\sigma-1)},$$

we can define the measure of buyers willing to pay at least p for quantity q as $1 - F(z(q, p))$. The seller chooses p to maximize per-unit profit:

$$\max_p \left[1 - F(z(q, p))\right](p - c).$$

The first-order condition yields:

$$\frac{p}{c} = \frac{\rho}{\rho - 1}, \quad \text{with } \rho = \frac{\sigma\kappa}{\sigma - 1},$$

which coincides exactly with the marginal price set under the optimal two-part tariff. The seller has no incentive to price discriminate beyond the optimal nonlinear schedule. This confirms both the optimality and robustness of the two-part tariff solution (see Appendix B.2 for details and a graphical explanation).

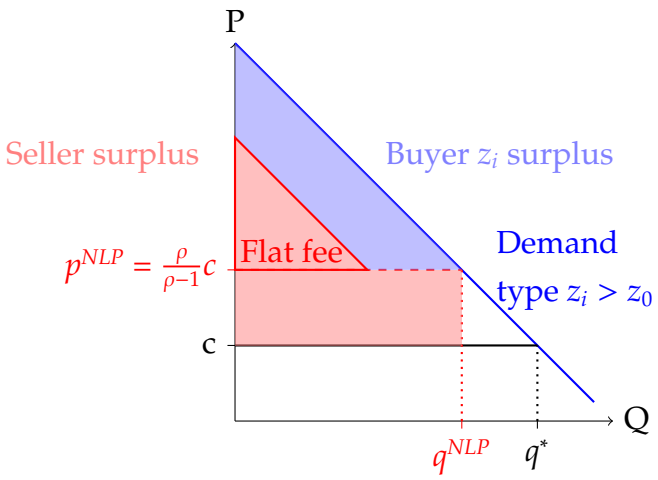
The optimal two-part tariff for a given buyer type $z_i > z_{\min}$ is illustrated in Figure 1.

[2]Virtual valuation represents the seller's marginal profit when accounting for both direct gains from serving a buyer type and the information rents required to prevent higher types from mimicking lower types. When virtual valuation is positive, it is optimal to serve every buyer type.

The seller sets a flat fee equal to the entire surplus that would accrue to the lowest type $z_{\min}$, thereby fully extracting their willingness to pay. In addition, the seller charges a constant per-unit price above marginal cost, marked up by a factor $\rho/(\rho - 1)$.

As a result, total seller surplus consists of two components: the flat fee (a lump-sum transfer) and the profits from the per-unit markup. The buyer's surplus under this scheme is reduced relative to a setting without a flat fee. Figure 1 illustrates these areas: the red-shaded regions represent seller surplus, while the blue-shaded region captures the residual buyer surplus for the higher type z_i. Figure 1 shows that the flat fee has no effect on the quantity allocated to buyer type z_i, which is completely determined by the marginal price p^{NLP}. We refer to this price as the allocative price component of the two-part tariff.

Figure 1: Optimal Two-Part Tariff Buyer Type z_i



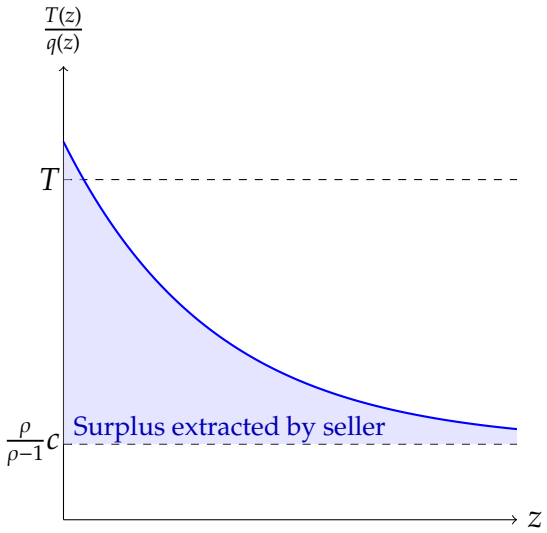
By dividing the total transfer $T(z) = F + pq(z)$ by the quantity purchased by type z, we can recover the average unit price paid by buyers:

$$\frac{T(z)}{q(z)} = \frac{F}{q(z)} + p$$

Figure 2 plots the average unit price $T(z)/q(z)$ as a function of buyer type z. The share of the flat fee in the average unit price decreases with buyer productivity: for low-productivity buyers, the average unit price is almost entirely driven by the flat fee, while for higher-productivity types a greater share of the total payment corresponds to the allocative component p.

The blue-shaded area represents the surplus extracted by the seller, which decreases monotonically with the buyer's type. This implies that quantity allocation to the lowest types is highly distorted relative to the efficient level, whereas allocation for the highest types is closer to efficient, with distortions diminishing as z increases.

Figure 2: Unit Average Price



Optimal Nonlinear Price Reproducibility. The optimal nonlinear pricing framework presented here is reproducible across a broad class of environments. Under a set of sufficient assumptions, the pricing logic extends naturally to more complex settings such as supply chains with multiple layers of buyers and sellers.

These sufficient conditions are: (i) revenue functions that are homothetic in quantity, of the form $R(q, z) = A(z)q^{\theta}$, where $A(z)$ is an arbitrary increasing function of buyer type and $\theta \in (0, 1)$; (ii) a Pareto distribution of buyer types with tail parameter $\kappa > \theta^{-1} - 1$, ensuring analytical tractability; and (iii) constant marginal cost of production, which may arise endogenously through competition or technology.

Under these conditions, a monopolistic seller facing buyers with private information about their types optimally offers a nonlinear price schedule in the form of a two-part tariff. The marginal (per-unit) price is constant across all types and given by a constant markup over marginal cost, $p = \frac{\rho}{\rho-1}c$, where $\rho = \theta\kappa/(1-\theta)$. The flat fee is designed to extract the entire surplus of the lowest type, binding the participation constraint. This contract structure remains unchanged regardless of the specific functional form of $A(z)$, the elasticity parameter θ, or the particular realization of buyer types, as long as revenue

remains homothetic.

This reproducibility becomes particularly powerful when extended to supply chains with multiple interconnected agents, where firms act as both buyers and sellers. At each layer of the chain, an upstream seller offers a two-part tariff to its downstream buyers, who in turn may resell using the same pricing logic.

Because the optimal nonlinear pricing rule is invariant in form, the structure of payments, incentives, and distortions remains consistent across all layers. This makes the framework well-suited for analyzing pricing distortions, efficiency losses, and surplus redistribution in supply chain environments where firm heterogeneity and market power in the form of nonlinear prices are prevalent.

2.2 Nonlinear Prices in Supply Chains

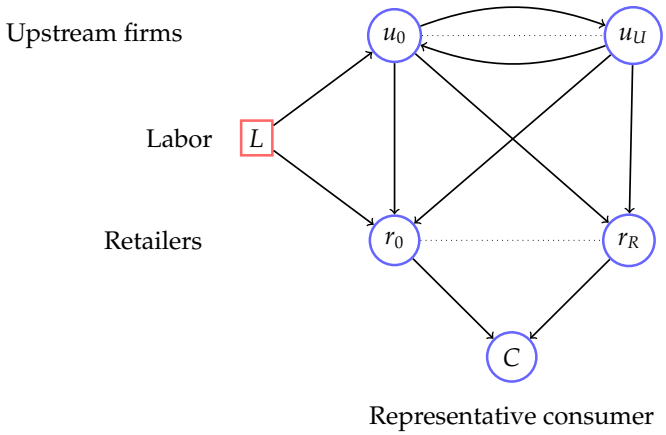
We develop a model of supply chains where firms implement nonlinear pricing strategies as sellers while facing nonlinear prices as buyers. We show that when firm-level productivities follow a Pareto distribution, profit-maximizing sellers optimally implement two-part tariffs. We provide the tools to examine the welfare implications of price discrimination compared to linear pricing and planner-optimal pricing, identifying the specific mechanisms that explain these differences.

Environment. There is a representative consumer who has preferences over final consumption goods, supplies labor as the only factor of production, and owns the firms. We model the supply chain as a two-layer structure consistent with a clear firm partition found in the data[3], where firms specialize in selling to other firms, which we call upstream firms, and retailers that sell primarily to final consumers.

In this framework, retailer firms sell exclusively to the representative household, while upstream firms both sell to and buy from other upstream firms but serve only as suppliers to retailer firms. There are $u \in U$ types of upstream firms and $r \in R$ types of retailers, where U and R denote the sets of possible types. Within each type, there is a continuum of firms. The economy is closed with no aggregate uncertainty, and the model is static. Figure 3 illustrates the production network structure of the economy.

[3]See Appendix A.5 for documentation of this fact.

Figure 3: Supply Chain Structure



Representative consumer. There is a representative consumer who derives utility from a constant elasticity of substitution (CES) aggregator over a continuum of goods produced by retailers:

$$Y = \left(\int_{r_0} y_r^{\frac{\sigma-1}{\sigma}} N_r \mu_r \, dr \right)^{\frac{\sigma}{\sigma-1}},$$

where σ represents the elasticity of substitution across retailer varieties, y_r denotes the consumption of retailer variety r, μ_r represents the density and N_r the mass of that variety. The representative consumer's utility is Y, which also corresponds to GDP in this economy.

The representative consumer inelastically supplies total labor L, which is allocated between production (L^P) and firm entry costs (L^E), so that $L = L^P + L^E$. The consumer receives wage income wL and owns both upstream firms and retailers. The consumer's budget constraint is:

$$YP_y = \Pi_U + \Pi_R + wL,$$

where P_y is the price index faced by the representative consumer and Π_U and Π_R are aggregate firm profits for upstream firms and retailers, respectively.

Production. Both upstream firms and retailers $i \in \{U+R\}$ produce using a Cobb-Douglas production function that aggregates labor and a bundle of intermediate goods sourced

from upstream firms:

$$q_i = z_i l_i^{\alpha} M_i^{1-\alpha},$$

where l_i represents labor input, M_i denotes the intermediate input bundle, and z_i is firm level productivity. The intermediate input bundle itself is a CES aggregator.[4]

We denote by m_{ui} the quantity purchased by firm i from upstream firm u. The intermediate input bundle is:

$$M_i = \left(\int_{u_0} m_{ui}^{\frac{\sigma-1}{\sigma}} N_u \mu_u \, du \right)^{\frac{\sigma}{\sigma-1}}$$

Firm entry and exit. To model firm entry, we rely on the frameworks of Hopenhayn (1992) and Melitz (2003) adapted to a supply chain structure. The model is static: firms decide to enter and, conditional on entry, produce. 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 a type from a layer-specific Pareto distribution G_U, G_R with shape parameters κ_U, κ_R. The masses of retailers N_R and upstream firms N_U are endogenous equilibrium objects.

2.3 Equilibrium Notions

We assume that firms are infinitesimal, so strategic interactions between firms are ruled out and each firm takes the prices of all other firms as given. We describe two different equilibrium notions. The first considers a setting where firms are constrained to offering only linear prices. In the second, upstream firms can offer different nonlinear prices to upstream buyers versus retailer buyers, based on observable buyer group characteristics.

In both equilibrium notions, retailers continue offering linear prices to the representative consumer. Moreover, in both cases, we impose feasibility constraints to ensure that supply equals demand. Additionally, with free entry, expected profits must equal the cost of entry for both upstream firms and retailers.

[4]We implicitly assume the production 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).

In Appendix B.3 we describe the standard feasibility and free entry conditions that apply to both pricing schemes, ensuring that supply equals demand and expected profits equal entry costs for entrants in both layers. We normalize the representative consumer price index to 1 ($P_y = 1$). We now illustrate the optimization problems that upstream firms and retailers face under monopolistic competition and charge linear prices.

2.4 Linear Pricing Equilibrium

The first equilibrium notion considers a setting in which all firms, upstream and retailers, are constrained to offering a single linear price under monopolistic competition. Firms demand inputs to minimize the cost of producing one unit of output:

$$\min_{\{m_{iu'}, l_i\}} \int_{u_0} p_{u'} m_{iu'} N_u \mu_{u'} du' + w l_i, \quad i \in \{R, U\}, \quad \text{s.t.} \quad q_i \geq 1,$$

which yields the following contingent input demand functions:

$$m_{iu'} = \left(\frac{p_{u'}}{p_m}\right)^{-\sigma} m_i,$$
$$m_i = \left(\frac{\alpha}{1-\alpha}\frac{p_m}{w}\right)^{-\sigma} l_i,$$

where p_m is given by:

$$p_m = \left(\int_{z_0} p(u)^{1-\sigma} \mu(u) N_u du\right)^{\frac{1}{1-\sigma}}.$$

Each firm's marginal cost is then given by:

$$mc_i = \frac{1}{z_i} c(w, p_m) = \frac{1}{z_i}\left(\frac{w}{1-\alpha}\right)^{1-\alpha}\left(\frac{p_m}{\alpha}\right)^{\alpha}.$$

Moreover, firms choose their prices to maximize total profits:

$$\max_{p_i} \left(p_i - mc_i\right) D_i, \quad i \in \{R, U\}$$

where D_i is total demand for firm type i. We assume that firms are infinitesimal and have no effect on other firms' pricing strategies, hence each firm sets its price as a constant

markup over marginal cost:

$$p_i = \frac{\sigma}{\sigma - 1} mc_i = \frac{\sigma}{\sigma - 1} \frac{1}{z_i} c(w, p_m).$$

Definition 1. *Decentralized Linear Pricing Equilibrium*
A decentralized linear pricing equilibrium is a collection of firm prices $\{p_r, p_u\}$, wage w, and quantities $\{y_r, q_r, l_r, m_{ru}, l_u, m_{uu'}, N_r, N_u\}$ such that, given technologies $\{z_r, z_u\}$: i) the representative consumer demands retailer goods to minimize costs, ii) firms 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, and v) feasibility constraints hold with equality.

Lemma 1. ***Efficient Benchmark***
The efficient benchmark is attained by a linear pricing decentralized equilibrium with an output subsidy equal to $\tau = \frac{\sigma-1}{\sigma}$ given to all firms in the economy and financed via a lump-sum tax on the representative consumer.

The lemma is a special case of the more general result in Theorem 1 of Baqaee and Farhi (2020). 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 equals expected profits, efficiency requires that expected profits equal 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 entry level.

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.

2.5 Nonlinear Pricing Equilibrium

Based on the basic construct of the optimal nonlinear price developed earlier, we now show the conditions needed to extend this framework to the supply chain model so that Proposition 1 can be implemented. In this setting, firms are no longer constrained to

offering a single linear price; instead, they can offer nonlinear pricing schedules to their buyers[5].

We assume that upstream firms can distinguish between upstream and retailer buyers, allowing them to offer different contracts to each buyer group. This constitutes a form of third-degree price discrimination based on observable buyer characteristics.

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 nonlinear prices set by their upstream suppliers. Additionally, as explained before, we use two additional assumptions. First, retailers charge a linear price to the representative consumer normalized to 1. Second, firms are infinitesimal and do not internalize other firms' outcomes or their actions' effects on other firms.

We extend the optimal nonlinear pricing solution to a multi-layer supply chain environment. Building on the benchmark monopolistic screening model, we show that the logic of two-part tariffs can be reproduced recursively along the supply chain under a set of sufficient conditions. These conditions ensure that firm behavior and contract design remain analytically tractable throughout the production network.

Sufficient Conditions for Recursive Optimal Pricing. To ensure that the optimal pricing logic extends to supply chains, we require three conditions. First, homothetic revenue functions: each buyer (upstream or retailer) generates revenue from intermediate inputs and labor according to a homothetic function in quantity and type:

$$R(q, z) = A(z)q^{\theta}, \quad \theta \in (0, 1),$$

where $A(z)$ is a strictly increasing function of the buyer's type z. This structure implies that the elasticity of revenue with respect to input quantity does not depend on scale, preserving the isoelasticity of marginal revenue curves.

In our model, this revenue form emerges endogenously from a Cobb-Douglas production function in labor and intermediates, combined with CES input demand:

$$y_i = z_i \cdot l_i^{1-\alpha} M_i^{\alpha},$$

where intermediates M_i are CES aggregates of upstream inputs. The dual revenue func-

[5]By allowing nonlinear pricing, we assume that resale markets are sufficiently costly to prevent arbitrage.

tion is then homothetic in effective inputs, satisfying the condition for recursive nonlinear pricing.

The second condition is Pareto distribution of firm types: both upstream firms and retailers have productivities z that follow a Pareto distribution

$$F(z) = 1 - \left(\frac{z_{\min}}{z}\right)^{\kappa}, \quad z \geq z_{\min}, \quad \kappa > 1.$$

This assumption implies a heavy-tailed productivity distribution and ensures analytical tractability in screening problems. Combined with homotheticity, it generates closed-form solutions for the optimal tariff schedule and participation decisions. The shape parameter κ governs the degree of heterogeneity and the distribution of surplus across firms, playing a role in welfare decompositions.

The third condition is constant marginal cost: each firm faces constant marginal cost, which is not imposed exogenously but follows from the equilibrium structure. Given CES technology in intermediate inputs and a fully connected supplier network, the unit cost function of firm i is:

$$c_i(w, p_m) = \frac{1}{z_i}\left[\left(\frac{w}{1-\alpha}\right)^{1-\alpha}\left(\frac{p_m}{\alpha}\right)^{\alpha}\right],$$

where w is the wage and p_m is the CES price index for intermediate inputs from upstream firms:

$$p_m = \left(\int_{u_0} p_u^{1-\sigma} N_u \mu_u \, du\right)^{\frac{1}{1-\sigma}}.$$

This implies that marginal cost is strictly decreasing in firm productivity z_i but does not vary with the scale of production. As a result, marginal cost remains constant across quantities for any given firm, which preserves the screening structure required for implementing nonlinear prices described in Proposition 1.

Under these conditions, the nonlinear pricing solution obtained in the monopolistic screening model can be extended recursively across the supply chain. Each seller, facing a continuum of heterogeneous buyers with private information about productivity, implements a menu of contracts that satisfies both incentive compatibility and individual rationality. The optimal contract takes the form of a two-part tariff:

$$T_{iu} = F_u(i_0) + p_u m_{iu}, \quad \text{for} \quad i \in \{R, U\},$$

where p_u is the marginal price charged by upstream firm u and $F_u(i_0)$ is the flat fee charged by upstream firm u that extracts all surplus from the lowest type in both buyer groups. By combining both components, the upstream seller screens buyer types.

Homotheticity ensures that the buyer's optimal quantity decision is isoelastic in their type i, and the Pareto distribution ensures that participation constraints can be solved in closed form. Constant marginal cost guarantees that the seller's profit maximization problem remains linear in output, preserving convexity and implementability of the schedule.

The recursive nature of this solution allows us to interpret the supply chain as a series of nested screening problems. Each upstream firm is both a buyer (of upstream inputs) and a seller (to downstream buyers), setting a two-part tariff that optimally extracts surplus from its buyers while taking upstream prices as given. The aggregation of these contracts across the supply chain yields the endogenous price index p_m, which feeds back into the cost structure of each firm, maintaining internal consistency.

As in the nonlinear price basic construct, it is always optimal for the seller to serve all buyer types and there is no profitable deviation by the seller to any buyer type.

This structure has implications for welfare analysis. Distortions arising from heterogeneous markups across buyer types and firm layers can be traced explicitly through the chain of two-part tariffs, allowing for clean decompositions of welfare losses into intensive and extensive margins, as developed in the following sections.

Definition 2. *Decentralized Nonlinear Pricing Equilibrium*
A decentralized nonlinear pricing equilibrium is a collection of linear prices $\{p_r, p_u\}$, flat fees $\{F_u(r_0), F_u(u_0)\}$, wage w, and quantities $\{y_r, q_r, l_r, m_{ru}, l_u, m_{uu'}, N_r, N_u\}$ such that, given technologies $\{z_r, z_u\}$: i) the representative consumer demands retailer goods to minimize costs, ii) both upstream firms and retailers 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 retailers and $\frac{\rho}{\rho-1}$ for upstream firms, v) every firm pays transfers such that the lowest types r_0, u_0 for retailers and upstream firms have zero surplus from transacting with upstream sellers, vi) entrants earn zero expected profit, and vii) feasibility constraints hold with equality.

Nonlinear Prices Implications. The total transfer follows a two-part tariff structure, consisting of a flat fee and a linear price component. The unit price paid by a firm can be computed by dividing the transfer by the quantity purchased. 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 buyers, 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 the linear pricing equilibrium, 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 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 uniform 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. The remaining portion of the markup, which does not distort quantity allocations, is a rent transfer from the buyer to the seller.

The allocative markup is lower than the markup in the linear pricing case, as $\rho > \sigma \Leftrightarrow \kappa > \sigma - 1$, which is a necessary condition for finite output.[6] 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, the planner would aim to eliminate this uniform tax to restore marginal cost pricing. Consequently, since the allocative markup is lower under the nonlinear 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.

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

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

equilibrium are given by:

$$\Pi_r = \frac{1}{\sigma - 1} c_r \left(\frac{z_r}{\tilde{z}_r}\right)^{\sigma} N_r^{\frac{\sigma}{1-\sigma}} Y, \quad \text{for retailers,}$$
$$\Pi_u = \frac{1}{\sigma - 1} c_u \left(\frac{z_u}{\tilde{z}_u}\right)^{\sigma} N_u^{\frac{\sigma}{1-\sigma}} D_m, \quad \text{for upstream firms,}$$

where $\tilde{z}_r, \tilde{z}_u$ are the average firm-level productivity for retailers and upstream firms, respectively. The profit functions under the nonlinear pricing equilibrium are given by:

$$\Pi_r = \frac{1}{\sigma - 1} c_r \left(\frac{z_r}{\tilde{z}_r}\right)^{\sigma} N_r^{\frac{\sigma}{1-\sigma}} Y - F_r N_u, \quad \text{for retailers,}$$

$$\Pi_u = \underbrace{\frac{1}{\rho - 1} c_u \left(\frac{z_u}{\tilde{z}_u}\right)^{\sigma} N_u^{\frac{\sigma}{1-\sigma}} D_m}_{\text{Profits from the linear price component}} + \underbrace{\left(\frac{p_u}{p_m}\right)^{1-\sigma} N_u (F_r N_r + F_u N_u)}_{\text{Collected flat fees}} - \underbrace{F_u N_u}_{\text{Flat fees paid}}, \quad \text{for upstream firms.}$$

For upstream firms, in expectation, the flat fees collected from and paid to other upstream firms cancel out exactly.[7]

Profits are distorted due to a redistribution of resources across firms relative to the linear pricing equilibrium: retailers pay flat fees to upstream firms. For retailers, the difference between the two profit functions manifests as a downward shift, since their variable profits remain unchanged while they pay flat fees to upstream firms.

For upstream firms, the distinction is ambiguous. First, the profits derived from the linear component of their two-part tariff are lower relative to linear prices, as they charge a reduced markup per unit. Additionally, they collect flat fees from all buyers while also paying flat fees to other upstream firms.

This implies that the wedge between profits in the linear and nonlinear pricing equilibria increases with firm productivity, with the most productive upstream 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.

[7] By definition $F_r = \mathbb{E}_u[F_{ru}]$ and $F_u = \mathbb{E}_u[F_{uu'}]$. Thus:

$$\mathbb{E}[\Pi_u] = \underbrace{\frac{1}{\rho - 1} c_u N_u^{\frac{\sigma}{1-\sigma}} D_m}_{\text{Profits from the linear price component}} + \underbrace{(F_r N_r + F_u N_u)}_{\text{Collected flat fees}} - \underbrace{F_u N_u}_{\text{Flat fees paid}}, \quad \text{for upstream firms.}$$

2.6 Inefficiencies

To describe how market power distortions affect aggregate welfare, we decompose total output and identify the inefficiencies that distort it relative to the planner's pricing equilibrium. To evaluate aggregate welfare under different pricing regimes, we compute total final output Y^{REG}, which aggregates firm-level production across the economy under a given regime $\text{REG} \in \{\text{LP}, \text{NLP}\}$.

Welfare in this setup depends on both the number of active firms (the extensive margin) and the average output per firm (the intensive margin). Let $\widehat{q}$ denote the weighted mean output per retailer:

$$\widehat{q} = \frac{1}{N_R} \left(\int_{r_0} q_r^{\frac{\sigma-1}{\sigma}} \cdot \mu_r N_r \, dr \right)^{\frac{\sigma}{\sigma-1}}$$

Aggregate welfare is given by:

$$Y^{\text{REG}} = \underbrace{(N_R^{\text{REG}})^{\frac{\sigma}{\sigma-1}}}_{\text{Extensive Margin}} \cdot \underbrace{\widehat{q}^{\text{REG}}}_{\text{Intensive Margin}}$$

This decomposition shows that welfare rises both with the number of active firms and with their average production scale. Under nonlinear pricing (NLP), firms internalize the marginal revenue distortions caused by linear pricing. The marginal price more closely aligns with marginal cost on inframarginal units, effectively reducing the double marginalization problem and inducing firms to produce more. This leads to higher average output per firm: $\widehat{q}^{\text{NLP}} > \widehat{q}^{\text{LP}}$

However, changes in the pricing regime also affect the composition of factor inputs. Under NLP, the higher effective price of intermediates (due to flat fees) induces firms to substitute away from intermediates M and toward labor L. Since incumbent firms operate at larger scale, they demand more labor for production, reducing the labor available for firm entry and leading to fewer entrants in the retail sector: $N_R^{\text{NLP}} < N_R^{\text{LP}}$

Whether NLP improves overall welfare relative to LP depends on which margin dominates. If the increase in intensive margin output outweighs the reduction in the mass of entrants, then total welfare improves. Conversely, if the contraction in entry is sufficiently large, LP may result in higher aggregate output.

The intensive margin can be decomposed further, starting with the fact that each firm's output is modeled as a Cobb-Douglas function of labor and material inputs. Using the

CES exponent:

$$q_r = z_r l_r^{\alpha} M_r^{1-\alpha}$$
$$q_r^{\frac{\sigma}{\sigma-1}} = z_r^{\frac{\sigma}{\sigma-1}} \left(\frac{l_r}{M_r}\right)^{\frac{\alpha\sigma}{\sigma-1}} M_r^{\frac{\sigma}{\sigma-1}}$$

Substituting the CES structure for material inputs, the CES-weighted average firm output is:

$$\begin{aligned}\widehat{q_R} &= \frac{1}{N_R}\left(\int_{r_0} z_r^{\frac{\sigma}{\sigma-1}} \left(\frac{l_r}{M_r}\right)^{\frac{\alpha\sigma}{\sigma-1}} \left[\int_{u_0} m_{ru}^{\frac{\sigma-1}{\sigma}} N_u \mu_u \, du\right]^{\frac{\sigma}{\sigma-1}} \mu_r N_r \, dr\right)^{\frac{\sigma-1}{\sigma}} \\ &= \bar{z_r}^{\frac{\sigma-1}{\sigma}} \left(\frac{\bar{l}_r}{\bar{M}_r}\right)^{\frac{\alpha\sigma}{\sigma-1}} \cdot \bar{M_r}^{\frac{\sigma-1}{\sigma}} \cdot N_U^{\frac{1}{\sigma}},\end{aligned}$$

where variables with a bar denote CES-weighted averages across downstream firms.

Thus, the ratio of CES-weighted output per firm between regime *REG* and the efficient benchmark *EFF* is:[8]

$$\frac{\widehat{q_R}^{REG}}{\widehat{q_R}^{EFF}} = \left(\frac{\bar{z_r}^{REG}}{\bar{z_r}^{EFF}}\right)^{\frac{\sigma-1}{\sigma}} \left(\frac{\bar{l_r}^{REG}/\bar{M_r}^{REG}}{\bar{l_r}^{EFF}/\bar{M_r}^{EFF}}\right)^{\frac{\alpha\sigma}{\sigma-1}} \left(\frac{\bar{M_r}^{REG}}{\bar{M_r}^{EFF}}\right)^{\frac{\sigma-1}{\sigma}} \left(\frac{N_U^{REG}}{N_U^{EFF}}\right)^{\frac{1}{\sigma}}$$

We can now express four distinct channels driving the intensive margin $\widehat{q^r}$. The first channel is retailer weighted average productivity ($\bar{z}_r$), which reflects how firm-level productivity translates into aggregate output under CES aggregation. Higher average productivity among downstream firms increases their effective contribution to output, conditional on input usage and allocation.

The second channel is input mix efficiency, captured by the average labor-to-material ratio $\bar{l}_r/\bar{M}_r$. This term indicates how far firms deviate from the cost-minimizing allocation of labor and materials. In the efficient regime, firms optimally allocate inputs to minimize costs. Under distorted pricing regimes, such as nonlinear pricing with markups on intermediates, materials become relatively more expensive, inducing substitution toward labor. This misallocation reduces output even when firm productivity remains unchanged, representing movement along a fixed isoquant driven by distorted input prices.

The third channel is the total quantity of material inputs used ($\bar{M}_r$). Even if firms

[8]Under the assumption that upstream input composition and firm distribution are sufficiently symmetric to simplify the inner integral over u to a function of N_U.

maintain the optimal input mix, they may still operate at inefficiently small scale due to elevated input prices or restricted access to materials. This reflects a scale distortion that shifts production from a higher to a lower isoquant, reducing aggregate output.

The fourth channel is upstream firm variety (N_U), which affects the efficiency of CES aggregation over differentiated intermediate inputs. A larger number of upstream firms lowers the CES price index for materials, improving input efficiency for downstream firms. Conversely, a reduction in upstream variety raises the effective price index and depresses downstream productivity. Together, these four channels explain how distortions in productivity, input mix, scale, and upstream variety jointly determine the intensive margin.

2.7 Optimal Policy

Having characterized the wedges that distort allocations under each pricing regime, we now derive the policy instruments required to implement the efficient allocation in decentralized equilibrium. As shown in Lemma 1, the efficient benchmark can be decentralized via a Linear Pricing (LP) equilibrium combined with an output subsidy that restores marginal cost pricing, conditional on entry.

Under LP, the constant markup imposed by retailers act as a tax on output, reducing both firm size and labor demand. Because firms are interconnected through intermediate input linkages, this production tax is amplified through input-output multipliers in the spirit of Jones (2011). In this setting, a single output subsidy—financed by a lump-sum tax on the representative consumer—fully restores efficiency.

The Nonlinear Pricing (NLP) equilibrium introduces additional distortions that impact both the intensive and extensive margins. On the intensive margin, four distinct mechanisms reduce average output per retailer.

First, allocative prices distort input choices by increasing the relative price of materials, leading firms to substitute toward labor and away from their cost-minimizing input mix—effectively moving along a fixed isoquant. Second, the flat-fee component of the two-part tariff extracts surplus from retailers, reducing their material usage and compressing firm scale. Third, upstream markups raise the effective price index of materials, limiting aggregate material usage and reducing upstream variety N_U, which further depresses retailer productivity.

Fourth, markup heterogeneity between sectors generates misallocation: retailers face the standard CES markup $\mu = \frac{\sigma}{\sigma-1}$, while upstream firms face a lower effective markup

due to the linear component of the two-part tariff. Although upstream markups are closer to efficient levels (since $\rho > \sigma$), this divergence from uniform marginal cost pricing introduces cross-sector inefficiencies.

These distortions are compounded by extensive margin effects. Flat fees charged to retailers reduce their expected profits and act as an implicit entry tax, distorting the mass of active retailers N_R. A similar logic applies to the upstream sector: distortionary markups reduce input demand, lowering expected profits and thus upstream entry N_U.

While upstream firms also pay flat fees to one another, these cancel out in expectation. Therefore, the entry distortion is asymmetric and takes the form of a net transfer from retailers to upstream firms. This inefficiency would not arise in a horizontally structured economy without vertical linkages, and can be neutralized by imposing a lump-sum tax on upstream firms, rebated to retailers.

Lemma 2. ***Optimal Policy under Nonlinear Pricing***
The efficient benchmark is achieved in a decentralized nonlinear pricing equilibrium with the following policy instruments:

- *Sector-specific output subsidies: $\tau_r = \frac{\sigma-1}{\sigma}$ for retailers and $\tau_u = \frac{\rho-1}{\rho}$ for upstream firms, financed via lump-sum taxes on the representative consumer.*
- *A lump-sum entry subsidy to retailers equal to the surplus extracted through flat fees, financed by a lump-sum entry tax on upstream firms.*
- *A lump-sum transfer to upstream firms to equate their expected profits with the marginal value of entry, financed via lump-sum taxes on the representative consumer.*

While output subsidies improve marginal incentives under NLP, the flat-fee-driven entry distortions introduce inefficiencies not present under LP. Whether the nonlinear scheme improves or worsens welfare relative to LP depends on the strength of these opposing forces: better allocative efficiency versus worse entry incentives.

Having established the theoretical framework, we now examine whether the predictions of our model are consistent with observed pricing patterns in the data. The theory predicts that profit-maximizing firms should implement nonlinear prices by buyer characteristics. In the following section, we test these predictions using the Chilean firm-to-firm transaction data, providing empirical validation for our theoretical assumptions before proceeding to quantitative welfare analysis.

3 Descriptive Evidence

Using Chilean Internal Revenue Service firm-to-firm transaction data for the year 2024, we examine price variations at the seller-product level. We find that nonlinear pricing is present, with quantity discounts (consistent with second-degree price discrimination) and pricing differences across buyer groups (consistent with third-degree price discrimination) accounting for 43% of observed price variation within seller-product pairs. We test whether the data generation process could be driven by buyer power rather than seller power and find patterns more consistent with seller price discrimination. These findings validate our modeling approach where sellers set nonlinear prices that vary by observable buyer characteristics.

Data Description. We use data from the universe of Chilean firm-to-firm value-added tax invoices collected by the Chilean Internal Revenue Service.[9] 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. These transaction records can be merged with firms' accounting variables, including total revenue, labor costs, materials costs, and capital expenditure.

We focus on the raw transaction data at its most granular level without excluding any industry. Our unit of observation is each line item within invoices between two tax identifiers.[10] In each line, we observe the transacted product "detail" which describes the product (for example, blue paint brand XX 3 gallons) and is often seller-specific; we refer to this variable as product. We observe each product's unit price and quantity.

While some aggregation and industry conditioning have proven helpful in uncovering certain characteristics, as Burstein et al. (2024) showed with the same database focusing on a subsample of manufacturing, we center our analysis on the most granular data available for the complete economy. In most transactions, shipping costs are recorded

[9]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 Chilean IRS. The information contained in the databases of the Chilean IRS is of a tax nature originating in self-declarations of taxpayers presented to the Service; therefore, the veracity of the data is not the responsibility of the Service

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

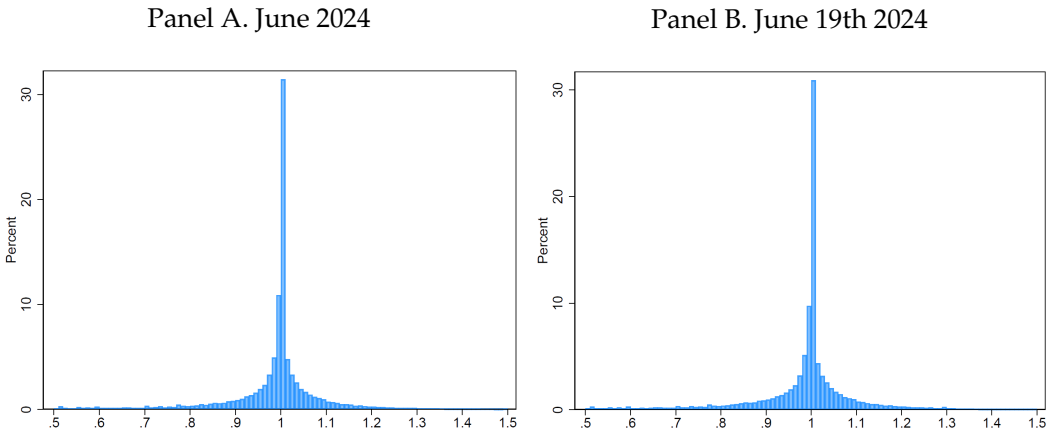
as a separate product on the invoice, implying that the unit prices in our analysis do not include shipping costs. One of the main challenges of the data is that we lack standardized information on product units, a limitation we address by treating products as firm-specific.

We perform three minimal data cleaning steps for consistency and to minimize capturing measurement error. First, we keep transactions with positive prices and quantities where the product detail variable is not missing or blank. Second, we keep firms that reported positive sales in at least one month during 2024. Third, to avoid measurement error in price variance, we drop products with at least 2 transactions where the price difference between the minimum and maximum observed price during a given day exceeds the 99th percentile of the maximum-to-minimum price ratio. This cleaning retains 98% of transactions. Our final sample contains 537,521 seller IDs and 3,398,323 buyer IDs that traded 60,029,741 different products across 1.24 billion transactions.

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 2024, the month with the most transactions in 2024. 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, 2024, the day with the most transactions in that month, to ensure our results are not driven by month-specific demand and supply shocks.

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 4. The histogram does not vary substantially from monthly to daily basis, while supply and demand shocks could still explain price differences. For 71% of transactions, we cannot reject that firms engage in some form of price discrimination departing from uniform pricing, although none of the exercises in this section aim to be causal, but rather describe equilibrium objects observed in the data and test which variables they correlate with in search of indicative evidence.

Figure 4: Price dispersion



Notes: The figure reports the distribution of the log of demeaned price for the month of June 2024. We exclude seller-product pairs with only one transaction.

Given this evidence that does not allow us to reject uniform pricing, we focus on identifying what could be the drivers departing from it consistent with price discrimination. While first-degree price discrimination, setting different nonlinear prices for each individual buyer based on their willingness to pay, is unlikely given the scale of transactions and information constraints faced by sellers in supply chains, 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 buyer groups).

In supply chains, it is empirically challenging to distinguish between second and third-degree price discrimination because prices typically vary along both the buyer and quantity dimensions. Second-degree discrimination arises when sellers offer nonlinear pricing menus (such as quantity discounts) and allow buyers to self-select, while third-degree discrimination involves offering different prices to different types of buyers based on observable characteristics. However, in practice, large or powerful buyers tend to purchase in greater quantities and also differ systematically from smaller buyers, making it difficult to determine whether lower unit prices are due to self-selection into quantity tiers or to targeted treatment based on buyer identity.

Moreover, sellers often use hybrid pricing strategies that combine nonlinear menus with buyer-specific negotiated terms, further blurring the distinction between the two mechanisms. Since researchers typically cannot observe counterfactual prices—such as what a different buyer would pay for the same quantity, or what the same buyer would

pay for a different quantity—it becomes difficult to cleanly attribute observed price variation to one form of discrimination over the other. As a result, disentangling second- and third-degree price discrimination in supply chains often requires strong assumptions or access to quasi-experimental variation that is rarely available in transaction-level data.

To explore the presence and structure of nonlinear pricing in supply chains, we proceed in two steps. First, we make transactions comparable by controlling for time-varying product-specific shocks. Specifically, we estimate the following regression:

$$\ln p_{igjt} = \beta_0 + \Psi_{igd} + \epsilon_{ijgt}, \tag{1}$$

where p_{igjt} is the transaction price between seller i, product g, buyer j, and time t. The fixed effect Ψ_{igd} absorbs all variation at the seller-product-day level, ensuring that remaining price differences cannot be attributed to common demand or supply shocks affecting all buyers of the same product from the same seller on the same day. The residuals ϵ_{ijgt} capture within-cell price variation that we then analyze in a second step. We regress these residuals on transaction-level observables, such as the quantity purchased and buyer characteristics, and compare the explanatory power of these variables using the resulting R^2 values. Combining buyer and quantity fixed effects is challenging because it requires multiple transactions within a day for the same seller-product-buyer combinations, which is unlikely. To address this, we define buyer groups as combinations of 11 sectors, 3 size categories, and 16 regions, yielding B distinct buyer types, making it possible to examine price variation from quantities and buyer observables simultaneously.

$$\epsilon_{ijgt} = \beta_0 + \Psi_{igdS} + \nu_{ijgt} \tag{2}$$

In the second step, we examine what drives the residual price variation ϵ_{ijgt} by regressing it on different combinations of quantity variation and buyer-side variation, represented by various fixed effect specifications Ψ_S. Given the empirical challenge of disentangling second- and third-degree price discrimination, we evaluate the explanatory power of quantity effects and buyer characteristics separately, and then jointly. The joint specification provides the most comprehensive assessment of price discrimination since it captures both quantity discounts and buyer group pricing simultaneously, reflecting the hybrid pricing strategies commonly used in supply chains.

Table 1 Column (1) shows that transaction quantity alone explains 26% of the residual

variation in prices, consistent with quantity discounts. Column (2) indicates that buyer fixed effects explain 31% of residual price variation, suggesting stable price heterogeneity across buyers. In Column (3), we replace buyer fixed effects with coarser buyer group fixed effects—defined by sector, size, and region—and still explain 23% of the variation, indicating that observable group characteristics capture a substantial portion of pricing patterns. Finally, Column (4) includes interaction terms between buyer group and quantity transacted, capturing both quantity discounts and group-specific pricing schedules. This specification explains 43% of the residual variation, suggesting that both forms of price discrimination are present and jointly relevant in supply chains. We run a similar specification using monthly fixed effects instead of daily fixed effects for manufacturing and retail and wholesale, the two largest industries by firm-to-firm transactions volume. We show the results in Appendix A.1 and find similar patterns of price determinants, consistent with nonlinear prices by buyer groups.

Table 1: Price residual determinants

	(1)	(2)	(3)	(4)
R^2	0.26	0.31	0.23	0.43
S = Quantity	✓			
S = Buyer		✓		
S = Buyer Group			✓	
S = Quantity × Buyer group				✓
N	147M	147M	147M	147M

Notes: The table reports R^2 values from regressions of price residuals ϵ_{ijgt} on different specifications S, where residuals are obtained from equation 2 after controlling for seller-product-day fixed effects. Buyer groups are defined by combinations of 11 sectors, 3 size categories, and 16 regions.

Nonlinear Prices. To test for the presence of nonlinear prices while remaining agnostic about the underlying data generation process, we examine whether equilibrium prices and quantities are systematically correlated. We estimate the following regression:

$$\ln p_{igjt} = \beta_0 + \beta_1 \ln q_{igjt} + \Psi_S + \epsilon_{ijgt}, \tag{3}$$

where p_{igjt} is the unit price charged by seller i for product g to buyer j on day d, and q_{igjt}

is the corresponding transaction quantity. The fixed effect Ψ_S varies across specifications and allows us to control for different combinations of seller, product, day, and buyer characteristics. All regressions are estimated using the universe of transactions in 2024, totaling over 430 million observations after dropping singletons.

Table 2: Price-quantity elasticity estimates

	(1)	(2)	(3)	(4)
$\ln q_{igjt}$	-0.042	-0.084	-0.065	-0.037
	(0.0001)	(0.0001)	(0.0001)	(0.0001)
S = Seller × Product × Day	✓			
S = Base + Buyer		✓		
S = Base + Buyer Group			✓	
S = Base × Buyer Group				✓
N	430M	430M	430M	430M
R^2	0.9646	0.9678	0.9659	0.9790

Notes: The table reports coefficients from regressions of log unit prices on log quantities with varying fixed effect specifications S. Base refers to seller × product × day fixed effects. Buyer groups are defined by combinations of 11 sectors, 3 size categories, and 16 regions. Standard errors in parentheses. All regressions use the universe of Chilean firm-to-firm transactions in 2024 after dropping singletons.

Column (1) reports the unconditional quantity coefficient, controlling only for seller-product-day fixed effects. The estimated coefficient of -0.042 implies that doubling the quantity purchased is associated with a 4.2% reduction in unit prices on average. In Column (2), we add buyer fixed effects. The quantity coefficient increases to -0.084, indicating that once we account for persistent buyer heterogeneity, quantity discounts become even more pronounced. Column (3) replaces buyer fixed effects with fixed effects for buyer groups—defined by sector, size, and region—and still yields a coefficient of -0.065.

Finally, Column (4) allows for flexible nonlinear pricing schedules across buyer groups by interacting the base fixed effects with buyer group. The quantity coefficient remains negative and statistically significant at -0.037, roughly 90% of the magnitude of the unconditional coefficient in Column (1).[11]

[11]We omit combining quantity fixed effects by buyer because of two reasons, there is not much sample where the same buyer and seller buy different quantities of the same product within the same day, and also

We repeat the same exercise from column (1) for each 1-digit sector in the economy and show the results in Appendix A.2. We find that the smallest quantity coefficient is around 0% in utilities while the largest is observed in construction at 13%. To analyze whether there are nonlinearities in the quantity discounts, we construct 50 quantiles for every product, where we modify the quantiles such that each quantity corresponds to the same quantile[12], and run the following regression:

$$\ln p_{ijgd} = \beta_0 + \sum_{b=1}^{50} \beta_b \mathbb{1}\{q_{ijgd} \in \text{quantile } b\} + \Psi_{igd} + \epsilon_{ijgd} \tag{4}$$

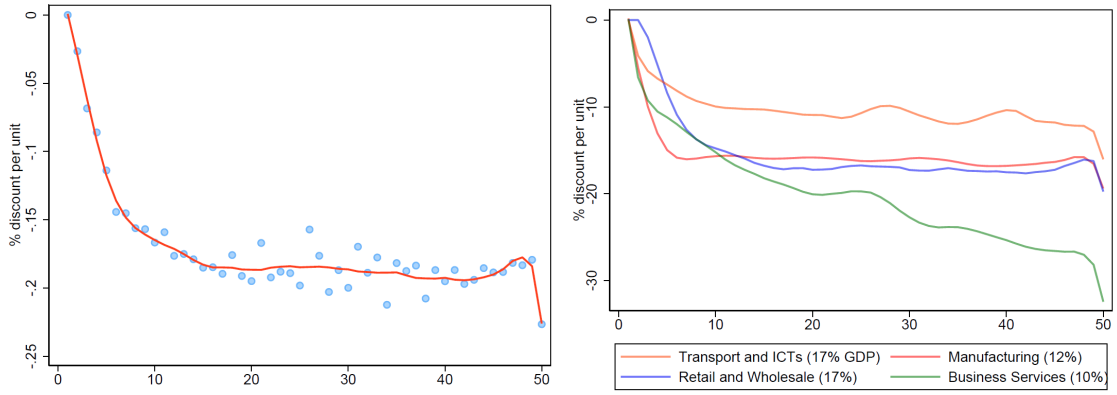
We estimate the coefficients for all quantity quantiles and plot them in Figure 5. Panel A shows that quantity discounts are steepest for small purchases, with discounts reaching approximately 15% by the 10th quantile, then stabilizing around 18-20% for larger quantities. This nonlinear pattern suggests that sellers use quantity discounts strategically to segment buyers and extract surplus from different transaction sizes. Panel B reveals substantial heterogeneity across industries in both the magnitude and shape of quantity discount schedules. Business Services exhibits the largest discounts, reaching 30% for high-quantity purchases, likely reflecting the scalable nature of service provision and lower marginal costs for large contracts. Manufacturing and Retail and Wholesale show more moderate discount schedules (15-18%), consistent with physical constraints and inventory costs that limit the ability to offer steep quantity discounts. Transport and ICTs display the flattest discount schedule, suggesting either more competitive pricing or cost structures that vary less with transaction size.

because doing so is evidence of nonlinear prices by buyer which is closer to first degree price discrimination which we assume to be implausible.

[12]We show the histogram of quantiles in Appendix A.4

Figure 5: Quantity Quantiles Bins Histogram

A. All B. Selected Seller Industries



Notes: The figure plots coefficients from regressions of log unit prices on quantity quantile indicators, controlling for seller-product-day fixed effects. Panel A shows results for all industries with a fitted trend line. Panel B shows results by major industry, with percentages indicating each industry's share of total firm-to-firm transaction volume. Quantity quantiles are constructed within each product to ensure comparable scales across heterogeneous products. The y-axis represents percentage discount per unit relative to the smallest quantity purchases. The x-axis shows quantity quantiles from 1 (smallest) to 50 (largest purchases).

An alternative explanation to seller-driven price discrimination is a data-generating process driven by buyer power. However, as shown in Table 1, including buyer fixed effects increases the magnitude of the estimated quantity coefficient from -0.042 to -0.084, which is more consistent with seller price discrimination than buyer bargaining power. If buyers were driving the quantity-price relationship through their bargaining strength, controlling for buyer heterogeneity should reduce rather than increase the quantity coefficient.

To further test the buyer power hypothesis, we construct a proxy for buyer power based on the number of suppliers a buyer transacts with. The underlying idea is that buyers with more suppliers have greater outside options and can more easily substitute across suppliers, potentially securing flatter quantity discount schedules. We estimate a regression interacting log quantity with this buyer power proxy and find that the coefficient on the interaction term is highly statistically significant but economically negligible, as detailed in Appendix A.3. This result indicates that buyers with access to more suppliers do not systematically receive different quantity discounts. Overall, the evidence suggests that buyer power is unlikely to be the main driver of the observed pricing patterns, which are more consistent with seller price discrimination.

Taken together, these results provide indicative evidence that quantity discounts are robust and systematically present in the data, even after accounting for different forms of buyer heterogeneity. This provides evidence in favor of rejecting uniform pricing and supports second-degree price discrimination as a central feature of firm-to-firm pricing behavior.

These findings are consistent with our theoretical model and motivate our stance on pricing conduct. We model sellers as engaging in price discrimination through nonlinear pricing schedules. To account for buyer relevance in pricing, we allow nonlinear prices to vary by buyer characteristics. Specifically, our model assumes that sellers can observe whether buyers are upstream firms or retailers and construct different nonlinear pricing schemes for each type. In the next section, we provide an additional test of nonlinear pricing by comparing the quantity discount schedule observed in the data with predictions from our model, using the Pareto tail of the buyer size distribution extracted from the data without targeting any pricing moments.

Having established that nonlinear pricing by buyer groups is present in Chilean supply chains, we now turn to bringing the model to the data to quantify the aggregate welfare implications of these pricing strategies. We calibrate our structural model to incorporate the empirical patterns documented above—quantity discounts and buyer group-specific pricing schedules—using the Chilean transaction data. This allows us to conduct counterfactual exercises comparing the observed nonlinear pricing equilibrium with alternative pricing regimes, including uniform pricing. The calibrated model provides a framework to assess how different pricing structures affect resource allocations and overall welfare in supply chains.

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 find that linear pricing regimes overestimate welfare costs relative to nonlinear pricing setups. The result is mainly driven by the intensive margin, specifically the expected output of retailers. Within the intensive margin, the main distortion explaining welfare losses relative to the planner's allocation is markup accumulation within the supply chain, creating a double marginalization problem.

4.1 Calibration

We calibrate the model using firm size distribution data and literature estimates, then validate the theoretical predictions against observed quantity discount patterns. The model successfully reproduces the nonlinear pricing patterns in the data without explicitly targeting them. The parameters include the labor share from input-output data and the Pareto tail parameter calibrated to match observed firm size distributions. The calibrated model captures both the shape and magnitude of quantity discounts, reaching approximately 24% for large purchases in both model and data.

Table 3 presents the parameter values used in our quantitative analysis. The labor share in production ($\alpha = 0.48$) is calibrated directly from input-output data, reflecting the proportion of total costs attributable to labor inputs. We assume that the elasticity of substitution across retailer varieties for the representative consumer equals the elasticity of substitution across upstream varieties for downstream firms. We set this parameter (σ) to 3 based on Hsieh and Klenow (2009).

The entry cost ($c_e = 0.0003$) is drawn from Hopenhayn et al. (2022). The Pareto tail parameter (κ) governs the productivity distribution and is calibrated to match the observed firm size distribution. In particular, the data reveal that firm size distribution measured as total number of workers follows a Pareto distribution with tail $\nu = 1.80$.[13] 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 $l(z) = l(\tilde{z})\left(\frac{z}{\tilde{z}}\right)^{\sigma-1}$. Thus, given the assumed σ and the estimated firm size distribution tail of ≈ 1.80, we calibrate κ accordingly.

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

Table 3: Model Parameters

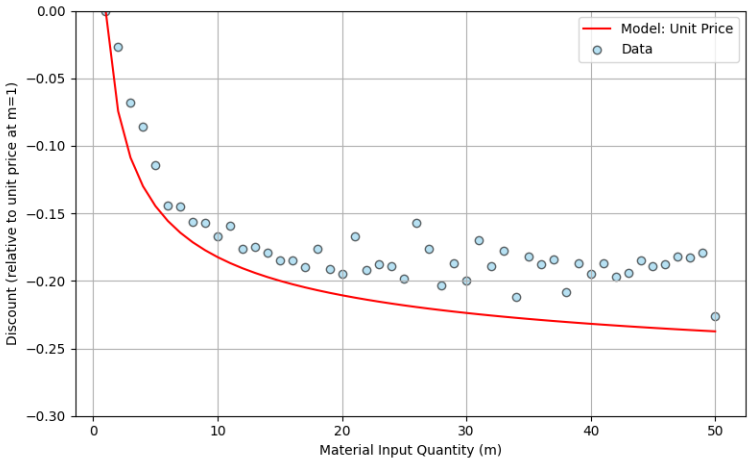
	Value	Source
Labor share in production (α)	0.48	Calibrated from data
Material bundle elasticity (σ)	3	Assumed
Entry cost (c_e)	0.0003	Assumed
Pareto tail (κ)	4.13	Calibrated from data

Notes: The table presents the four parameters used in the quantitative analysis. The labor share is calibrated from Chilean input-output data. The elasticity of substitution is set to 3 following Hsieh and Klenow (2009). Entry cost is drawn from Hopenhayn et al. (2022). The Pareto tail parameter is calibrated to match the observed firm size distribution tail of 1.80.

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 observed quantity discounts using the empirical strategy from the descriptive evidence section. The figure illustrates the unit prices charged by the average upstream firm to retailers as a function of the quantity purchased, corresponding to productivity values z ranging from z_0 to the 99th percentile of the z distribution, with quantities normalized to range from 1 to 50.

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 negative relationship between unit prices and purchase quantities, where larger purchases receive larger quantity discounts. The model also captures quantity discount levels, which reach approximately 24% in both the model and the data.

Figure 6: Calibrated Model Unit Prices



Notes: The figure compares model-predicted quantity discounts with observed patterns in the data. Model unit quantity discounts are computed for the average upstream firm price schedule to retailers, normalizing the continuous input quantity to range from 1 to 50. The model reproduces both the shape and magnitude of quantity discounts without explicitly targeting these patterns, reaching approximately 24% for large purchases in both model and data.

4.2 Quantification

We quantify welfare losses by decomposing aggregate output into intensive and extensive margins, finding that nonlinear pricing achieves 66% of efficient welfare compared to 59% under linear pricing. Material usage distortions dominate welfare losses, accounting for 65-71% of intensive margin inefficiencies through double marginalization effects. While both regimes generate excess entry, nonlinear pricing performs better on total firm mass but creates more severe upstream firm shortages. The intensive margin accounts for roughly 63% of total welfare losses, with extensive margin effects contributing 37%.

We now quantify the welfare impact of linear and nonlinear pricing regimes relative to the efficient benchmark by decomposing losses into intensive and extensive margin components. We use the exact CES formula for aggregate output:

$$Y^{REG} = \left(N_R^{REG}\right)^{\sigma/(\sigma-1)} \cdot \widehat{q}^{REG}$$

where $\widehat{q}^{REG}$ denotes the CES-weighted average output per retailer. As derived earlier, this intensive margin can be decomposed into four interpretable components: firm productivity selection, input mix efficiency, material usage, and upstream input variety.

Table 4 shows that productivity selection is unaffected across regimes, reflecting identical entry thresholds. However, both distorted regimes exhibit inefficiencies along the other three channels. Input mix distortions are more severe under LP due to uniform markups on all inputs, while NLP improves marginal input allocation by applying lower markups to upstream trades. However, NLP introduces scale distortions due to flat fees that reduce material usage. This effect, captured in the Material usage row, is severe in both regimes but slightly less under NLP. Finally, NLP reduces upstream variety (N_U) more than LP, reflecting stronger disincentives for upstream firm entry due to the flat fee structure.

Table 4: Intensive Margin Components Relative to Efficient Benchmark

Intensive Margin Component	**Linear Pricing (LP)**	**Nonlinear Pricing (NLP)**
Productivity Selection ($\bar{z}_r$)	1.00	1.00
Input Mix ($\bar{l}_r/\bar{M}_r$)	1.36	1.29
Material Usage ($\bar{M}_r$)	0.41	0.48
Input Variety (N_U)	0.94	0.87

Notes: The table decomposes the intensive margin into four components relative to the efficient benchmark (=1.00). Productivity selection reflects the average productivity of active retailers and is unaffected across regimes. Input mix captures the labor-to-materials ratio, with values above 1 indicating excessive substitution toward labor due to material markups. Material usage shows the scale of material inputs, with values below 1 indicating undersupply due to markups and flat fees. Input variety reflects the number of upstream firms available as input suppliers.

Turning to the extensive margin, in Table 5 both regimes increase the mass of active firms relative to the planner. However, the distortions are asymmetric. Both regimes result in too many retailers and too few upstream firms relative to the efficient allocation. NLP exacerbates the upstream shortage more severely than LP, with upstream firms falling to 66% of the efficient level under NLP compared to 85% under LP. Similarly, NLP reduces retailer entry compared to LP, though both regimes still have excess retailers. Overall, NLP is closer to the planner in total firm mass and thus performs better on this margin.

Table 5: Extensive Margin: Mass of Active Firms Relative to Efficient Benchmark

Firm Mass Component	**Linear Pricing (LP)**	**Nonlinear Pricing (NLP)**
Total Firm Mass (N)	1.23	1.07
Upstream Firms (N_U)	0.85	0.66
Retailers (N_R)	1.65	1.51

Notes: The table shows the mass of active firms relative to the efficient benchmark (=1.00). Both regimes generate excess total entry (values above 1) but with asymmetric distortions across layers. Values above 1 indicate too many firms, while values below 1 indicate too few firms. Both regimes result in too many retailers and too few upstream firms, with NLP exacerbating the upstream shortage while moving total firm mass closer to efficient levels.

Both pricing regimes generate excess entry relative to the planner because markup profits create entry incentives that exceed social value. In the efficient benchmark, entry should occur until the social value of an additional firm, including both producer surplus and consumer surplus from variety, equals the entry cost. However, in the decentralized equilibria, firms base entry decisions solely on private profits. Markup profits, particularly for retailers selling to consumers, make entry more attractive than socially optimal. This reflects the tension in CES economies with love-of-variety preferences: while variety is socially valuable, markup distortions create private rents that exceed the social contribution of marginal entrants.

Welfare under NLP is 66% of the planner's benchmark, compared to 59% under LP as shown in Table 6. This implies that NLP eliminates roughly 18% of the welfare loss observed under LP, improving welfare from 59% to 66% of the efficient level. When decomposing the welfare gap in logs, we find that roughly 63% of the deviation arises from intensive margin distortions, and 37% from extensive margin distortions. Even with free entry, misallocation across and within firms, especially via material usage and input mix distortions, dominates the welfare loss.

Table 6: Welfare Decomposition Relative to Efficient Benchmark

Measure	**LP**	**NLP**
Welfare (Y^{REG}/Y^{EFF})	0.59	0.66
Intensive Margin ($\widehat{q}^{REG}/\widehat{q}^{EFF}$)	0.42	0.50
Extensive Margin (N_R^{REG}/N_R^{EFF})	$1.65^{\sigma/(\sigma-1)}$	$1.50^{\sigma/(\sigma-1)}$
Intensive Margin (Log Share of Gap)	63%	62%
Extensive Margin (Log Share of Gap)	37%	38%

Notes: The table decomposes welfare relative to the efficient benchmark. Total welfare equals the intensive margin (average output per retailer) times the extensive margin (number of retailers) raised to the CES power. NLP achieves 66% of efficient welfare compared to 59% under LP. Log shares show the contribution of each margin to the total welfare gap, calculated as the log deviation from efficient levels. The intensive margin dominates welfare losses under both regimes.

Table 7 makes clear that among intensive margin distortions, inefficient material usage—driven by upstream markups and flat fees—accounts for the majority of the welfare gap under both pricing regimes, representing 71% under LP and 65% under NLP. This double marginalization channel is the dominant quantitative source of inefficiency.

Table 7: Log Deviation of Intensive Margin Components from Planner

Component	**LP (Log Share)**	**NLP (Log Share)**
Productivity Selection	0%	0%
Input Mix	25%	23%
Material Usage	71%	65%
Input Variety	4%	12%

Notes: The table shows the contribution of each intensive margin component to the total intensive margin welfare gap in log terms. Log shares are calculated as the log deviation of each component from efficient levels, expressed as a percentage of the total intensive margin log deviation. Material usage dominates welfare losses through double marginalization effects, accounting for 65-71% of intensive margin inefficiencies. Input mix distortions contribute 23-25% through substitution away from materials due to markups.

To further isolate intensive distortions, we perform a counterfactual where the total number of firms is fixed at the efficient level, but allowed to reallocate endogenously

across layers. Table 8 shows that even when $N^{REG} = N^{EFF}$, welfare under NLP remains at 0.64, only marginally below its baseline level of 0.66. This confirms that distortions to material usage, input mix, and reduced upstream variety remain first-order even when extensive margin distortions are eliminated.

Table 8: Fixing Total Firm Mass: Intensive Margin Under NLP

Component	**Fixed Entry (NLP)**	**Baseline NLP**
Productivity Selection	1.00	1.00
Input Mix	1.27	1.29
Material Usage	0.49	0.48
Input Variety	0.86	0.87
Welfare	0.64	0.66

Notes: The table shows a counterfactual where total firm mass is fixed at the efficient level but firms can reallocate endogenously across upstream and retail layers. Fixed Entry (NLP) shows results under this constraint, while Baseline NLP shows the unconstrained equilibrium. All components are relative to the efficient benchmark (=1.00). Welfare declines only marginally from 0.66 to 0.64, confirming that intensive margin distortions dominate welfare losses even when extensive margin distortions are eliminated.

The quantitative analysis reveals that nonlinear pricing provides welfare improvements over linear pricing, achieving 66% of efficient welfare compared to 59% under linear pricing. This 18% reduction in welfare losses occurs despite more complex distortions under nonlinear pricing, including asymmetric entry effects and heterogeneous markups across firm types. The dominance of intensive margin effects—accounting for roughly 63% of welfare losses—highlights that misallocation within existing firms, particularly through material usage distortions and double marginalization, represents the primary source of inefficiency in both regimes. While both pricing structures generate excess entry relative to the planner, the model's ability to reproduce observed quantity discount patterns without explicitly targeting them provides confidence in these welfare calculations. The results suggest that policies aimed at reducing markup accumulation along supply chains may yield larger welfare gains than those focused solely on correcting entry distortions.

5 Conclusion

This paper studies the macroeconomic effects of price discrimination in firm-to-firm transactions, focusing on how it shapes firm entry, resource allocation, and aggregate welfare. Using granular transaction-level data from the universe of Chilean firms, we document the widespread use of both second- and third-degree price discrimination in supply chains.

We develop a supply chain general equilibrium model with nonlinear pricing and show that, relative to linear pricing, price discrimination reduces allocative distortions but compresses firm profits, especially for marginal entrants. Despite these entry effects, nonlinear pricing generates higher welfare, suggesting that standard models with linear prices overstate the welfare costs of market power.

Our findings raise policy implications: if market power is present, banning price discrimination may backfire. In our setting, allowing firms to discriminate across buyers improves allocations and raises welfare, even though entry is reduced. Regulatory frameworks that restrict pricing flexibility in the presence of market power should weigh these trade-offs carefully.

This work opens several avenues for future research. First, incorporating endogenous production networks could reveal how firm-to-firm linkages evolve under different pricing regimes. Second, understanding how pricing strategies respond to shocks—such as cost or demand shifts—can inform countercyclical policy. Finally, richer empirical tests of market conduct and pricing at finer levels of buyer–seller granularity would enhance identification of pricing power mechanisms.

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.

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.

Varian, H. R. (1989). Price discrimination. *Handbook of industrial organization*, 1:597–654.

Wilson, R. B. (1993). *Nonlinear pricing*. Oxford University Press, USA.

Appendix

A Additional Descriptive Evidence

A.1 Residual Price Determinants by Selected Industries

To assess whether the pattern of nonlinear pricing driven by buyer observables generalizes across sectors, we replicate the residual decomposition analysis for the two industries

with the highest volume of transactions: Manufacturing and Retail and Wholesale. For both sectors, we estimate the following regression:

$$\ln p_{igjt} = \beta_0 + \Psi_{igdS} + \epsilon_{ijgt}, \tag{5}$$

where p_{igjt} is the unit price of a product g sold by seller i to buyer j at time t, and Ψ_{igdS} represents seller-product-day fixed effects interacted with different sets of buyer-side controls S. Buyer groups B are defined based on 11 sectors, 3 firm-size categories, and 16 regions.

Table 9: Price residual determinants: Manufacturing

	(1)	(2)	(3)	(4)
Adjusted R^2	0.45	0.54	0.46	0.81
S = Quantity	✓			
S = Buyer		✓		
S = Buyer Group			✓	
S = Quantity × Buyer Group				✓
N	136M	136M	136M	136M

Table 10: Price residual determinants: Retail and Wholesale

	(1)	(2)	(3)	(4)
Adjusted R^2	0.38	0.65	0.49	0.68
S = Quantity	✓			
S = Buyer		✓		
S = Buyer Group			✓	
S = Quantity × Buyer Group				✓
N	180M	180M	180M	180M

In both sectors, the pattern remains unchanged: quantity discounts (second-degree price discrimination) and buyer group-based pricing (third-degree) explain the majority

of price dispersion once product and time effects are controlled for. This reinforces our main finding that nonlinear prices shaped by buyer-side observables are a pervasive feature of pricing in supply chains.

A.2 Average Quantity Discount by Sector

Table 11: Average Quantity Discount by Sector

Sector	**Mean Q discount**	**N transactions**
All sectors	-0.042	430M
Agriculture	-0.042	2M
Mining	-0.016	1M
Manufacturing	-0.036	118M
Utilities	0.000	6
Construction	-0.129	1M
Retail and Wholesale	-0.048	270M
Transport & ICTs	-0.032	12M
Financial Services	-0.002	49M
Real Estate Services	-0.052	1M
Business Services	-0.089	5M
Personal Services	-0.053	1M

A.3 Test for Buyer Power Data Generation Process

To examine whether observed quantity discounts reflect buyer power rather than seller-driven price discrimination, we exploit cross-sectional variation in the number of suppliers each buyer transacts with during the sample period. The underlying idea is that buyers with access to a larger number of sellers may possess stronger outside options, enhancing their bargaining position and enabling them to negotiate better pricing terms. We define buyer power as the logarithm of the total number of distinct sellers each buyer purchases from within the observed month. We then test whether buyer power flattens quantity discounts by estimating the interaction between log quantity and buyer power

in a log-linear price regression. Specifically, we estimate:[14]

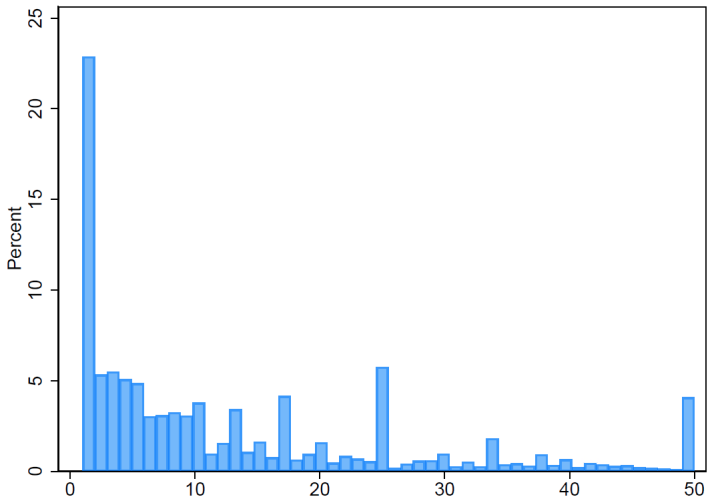
$$\ln p_{igjt} = \beta_0 + \beta_1 \ln q_{igjt} + \beta_2 \left(\log q_{ijg} \times \log \text{NumProviders}_j\right) \Psi_{igd} + \epsilon_{ijgt},$$

A positive coefficient on the interaction term ($\beta_2 > 0$) would suggest that quantity discounts become flatter as buyer power increases, consistent with buyers using their broader supplier base to resist steep discounts or nonlinear price schedules.

We find that $\beta_1 = -0.0462$ and $\beta_2 = -0.0098$, both estimated with standard errors below 0.0001. While the interaction term is statistically significant, the magnitude is economically negligible. This suggests that buyer power, as measured by the number of suppliers, does not appear to be the primary mechanism generating quantity discounts. If anything, the evidence is more consistent with seller-driven price discrimination rather than buyer power shaping quantity discounts.

A.4 Quantity Discount Quantiles Bins

Figure 7: Quantity Quantiles Bins Histogram



[14]Standard errors are clustered at the buyer level to account for within-buyer correlation.

A.5 Firm Sales Partition

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 12: 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 12, there is heterogeneity across sectors, though the partition between firms selling to final consumers and other firms is present across all sectors.

B Model Details

B.1 Virtual Surplus

The monopolist optimally chooses to serve all buyer types, including the lowest type $z_0 = 1$. This follows from the analysis of virtual surplus and the structure of the buyer type distribution. The monopolist faces a standard trade-off in nonlinear pricing with private information: Exclude low types to better extract surplus from high types (i.e., leave less informational or serve all types but give up some surplus from high types due to binding IC constraints.

However, when buyer types follow a Pareto distribution, the density of low types is large, which solves the trade-off. Even though each low type contributes little output, their large mass makes serving them profitable.

We showed that the the seller's pointwise virtual surplus from serving a buyer type z is:

$$\text{VS}(z) = \left(\frac{z^{\sigma-1}}{\alpha} - \frac{1 - F(z)}{f(z)} \cdot \frac{d}{dz} \left(\frac{z^{\sigma-1}}{\alpha} \right) \right) q(z)^{1-1/\sigma} - cq(z).$$

This expression captures, the marginal gain from selling to type z: $\frac{z^{\sigma-1}}{\alpha} q(z)^{1-1/\sigma}$ and also the informational rent that must be left to higher types to maintain IC, subtracting $cq(z)$ accounts for production cost. Substitute $z = 1$ into the virtual surplus expression. Recall:

$$\frac{1 - F(z)}{f(z)} = \frac{z}{\kappa}, \quad \text{and} \quad \frac{d}{dz} \left(\frac{z^{\sigma-1}}{\alpha} \right) = \frac{(\sigma - 1) z^{\sigma-2}}{\alpha}.$$

So:

$$\begin{aligned} \text{VS}(z = 1) &= \left(\frac{1}{\alpha} - \frac{1}{\kappa} \cdot \frac{(\sigma - 1)}{\alpha} \right) q(1)^{1-1/\sigma} - cq(1) \\ &= \frac{1}{\alpha} \left(1 - \frac{\sigma - 1}{\kappa} \right) q(1)^{1-1/\sigma} - cq(1). \end{aligned}$$

The first term is strictly positive if $\kappa > \sigma - 1$. This is exactly the same condition that ensures expected output in the model, and is assumed throughout. Hence, for small enough $q(1)$, the gain exceeds the cost: $\text{VS}(1) > 0$.

Since the virtual surplus from serving the lowest type is strictly positive, and the density of low types is large (Pareto distribution), excluding them reduces total profits. Thus

the seller optimally chooses to serve all buyer types. In screening models with heavy-tailed distributions, the cost of excluding low types outweighs the gains from extracting additional surplus from high types.

B.2 No Profitable Price Deviation

We show that the seller does not benefit from deviating from the optimal nonlinear pricing schedule by charging different per-unit prices for the same quantity.

Buyers solve:

$$\max_q \left\{ \frac{z^{(\sigma-1)/\sigma} q^{(\sigma-1)/\sigma}}{(\sigma-1)/\sigma} - T(q) \right\},$$

which implies the first-order condition:

$$z^{(\sigma-1)/\sigma} q^{-1/\sigma} = T'(q) = p(q).$$

This condition yields an inverse demand curve for the q^{th} unit:

$$z(q,p) = q^{1/(\sigma-1)} p^{\sigma/(\sigma-1)}.$$

The seller could try to deviate and charge a different price p for a given q. The total demand for that unit is:

$$D(q,p) = 1 - F(z(q,p)),$$

and the profit from this price is:

$$\max_p \left[1 - F(z(q,p))\right](p-c).$$

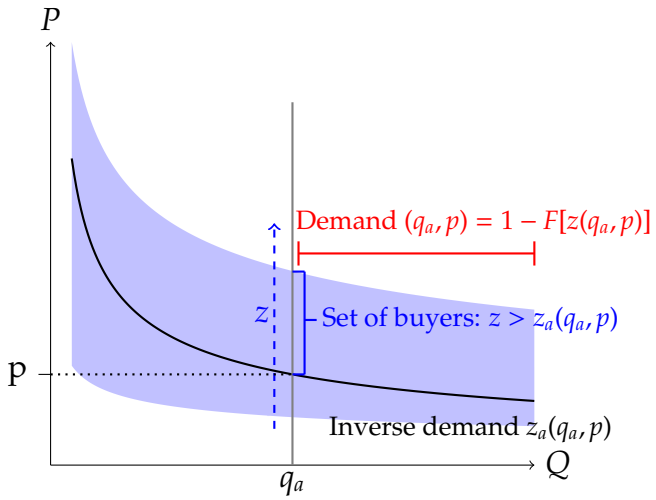
Using the Pareto distribution $F(z) = 1 - z^{-\kappa}$, we get:

$$\frac{d}{dp}\left[z(q,p)^{-\kappa}(p-c)\right] = 0.$$

Solving yields the optimal price:

$$\frac{p}{c} = \frac{\rho}{\rho - 1}, \quad \text{where } \rho = \frac{\sigma\kappa}{\sigma - 1}.$$

Figure 8: No Profitable Price Deviation



This is exactly the same price as that used in the optimal two-part tariff. Thus, the price that maximizes marginal profit from each q-unit is identical to the allocative price already embedded in the mechanism. There is no gain from deviating to another price scheme for a subset of buyers. Figure 8 illustrates this intuition: even if a seller attempts to increase p for all buyers that buy q_a for above, the marginal buyers drop out, and the loss in volume offsets any price gain, such that the optiminal nonlinear price remains unchanged. The nonlinear price schedule constructed from the mechanism design problem is fully incentive compatible, making the solution robust to an euristic approach using Wilson (1993) logic.

B.3 Feasibility and Zero Profit Condition

Feasibility. For every good produced, whether upstream or by retailers, demand does not exceed supply. Thus, for retailers' goods, feasibility requires:

$$y_r \leq q_r \quad \forall r,$$

where y_r is the representative consumer quantity demanded for retailer r, and q_r is the quantity produced by retailer z. Similarly, for upstream goods:

$$\int_{r_0} m_{ru}\mu_r N_r dr + \int_{u_0} m_{u'u}\mu'_u N_{u'} du' \leq q_u \quad \forall u.$$

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:

$$\int_{r_0} l_r \mu_r N_r dr + \int_{u_0} l_u \mu_u N_u du + c_e(N_R^E + N_U^E) \leq 1,$$

where N_R^E and N_U^E denote the mass of entrants in the retailers and upstream firms, respectively. Furthermore, let L^P denote the total amount of labor used in production by upstream firms and retailers:

$$L^P = \int_{r_0} l_r \mu_r M_r dr + \int_{u_0} l_u \mu_u N_u du.$$

Substituting this condition into the labor feasibility constraint, labor can be used either in production or to create firms, so that we obtain:

$$L^P + c_e(N_R^E + N_U^E) \leq 1.$$

Zero Profit Condition. In an equilibrium with unrestricted entry, expected profits must equal the cost of entry in both layers. Let π denote firm level profits, then:

$$\begin{aligned} \mathbb{E}_r[\pi_r] &= c_e w, \quad \text{For retailers} \\ \mathbb{E}_u[\pi_u] &= c_e w, \quad \text{For upstream firms} \end{aligned}$$

B.4 Optimal Nonlinear Price in Supply Chains

We assume that firm-level productivity follows a Pareto distribution with tail parameter κ. For output to be finite, we require the tail condition $\kappa > \sigma - 1$[15] . As shown in the

[15]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

descriptive evidence, this assumption generates nonlinear pricing patterns.

We proceed by applying the method of guessing and verifying.

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

$$T_{iu} = F_u(i_0) + p_u m_{iu}, \quad \text{for} \quad i \in \{R, U\}.$$

Guess 2: The revenue functions take the following form:

$$R_i = q_i^{\theta} A_i, \quad \text{for some} \quad \theta, A_i, \quad i \in \{R, U\}.$$

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

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

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

$$p_m = \left(\int p_u^{1-\sigma} N_u \mu_u \, du \right)^{\frac{1}{1-\sigma}}.$$

Moreover, the pricing problem of retailers, 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 $\{T_{iu}, m_{iu}\}$, where T_{iu} denotes the total transfer from firm type i to the upstream seller u. 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 Π_i the total profits of a buyer-firm of type and define the total buyer surplus from purchasing quantity x from seller u as:

$$\tilde{\Pi}(i, x) = \frac{d\Pi_i}{d\left(\mu_u N_u\right)}\Big|_{m_{iu}=x}$$

finite average, $\frac{\kappa}{\sigma-1} > 1$

The upstream seller's problem is:

$$\max_{\substack{\{T_{ru},m_{ru}\},\\ \{T_{u'u},m_{u'u}\}}} \Pi_u = \underbrace{\mathbb{E}_r\left[T_{ru} - c(z_u)\right]}_{\text{Profits from retailers}} + \underbrace{\mathbb{E}_{u'}\left[T(u',u) - c(z_u)\right]}_{\text{Profits from other upstream firms}}$$

subject to:

$$\begin{aligned} &\text{(IR)} \quad \tilde{\Pi}(u, m_{iu}) \geq 0 \quad \forall \quad i \in \{R, U\} \\ &\text{(IC)} \quad m_{iu} \in \underset{m_{iu}}{\operatorname{argmax}}\, \tilde{\Pi}(u, m_{iu}) \quad \forall \quad i \in \{R, U\}, \end{aligned}$$

where $\tilde{\Pi}(u, m_{iu})$ is the surplus of firm i of the transaction with upstream firm u. Under the proposed guesses, and using buyers profit expressions, the upstream firm's problem can be written as the sum across buyers layers i where i can be upstream buyers or retailers, and within each layer the integral of all layer types:[16]

$$\max_{\{m(\tau_{iu}),\tilde{\Pi}(\tau_{iu})\}} \Pi_u = \sum_{i\in R,U}\left(\int_{\tau(z_{i=0})}\left[(\tau_{iu} - h^{-1}(\tau_{iu}))m(\tau_{iu})^{\frac{\sigma-1}{\sigma}} - c_u m(\tau_{iu}) - \tilde{\Pi}(\tau(z_{i=0}))\right]\tau_{iu}N_i\right),$$

with:

$$\tau_{iu} = A_i z_i^{\theta}\theta y_i^{\theta-1}\left[\alpha l_i^{\frac{\sigma-1}{\sigma}} + (1-\alpha)m_i^{\frac{\sigma-1}{\sigma}}\right]^{\frac{1}{\sigma-1}}(1-\alpha)m_i^{-\frac{1}{\sigma}}\frac{\sigma}{\sigma-1}m_i^{\frac{1}{\sigma}}.$$

Where τ_{iu} represents the buyer's heterogeneous valuation for the good sold by the seller, a strictly increasing function of buyer level productivity z_i. τ_{iu} is distributed according to a Pareto distribution with tail parameter $\rho := \frac{\kappa\sigma}{\sigma-1}$.

In particular, under the guess, notice that the term:

$$A_i z_i^{\theta}\theta y_i^{\theta-1}\left[\alpha l_i^{\frac{\sigma-1}{\sigma}} + (1-\alpha)m_i^{\frac{\sigma-1}{\sigma}}\right]^{\frac{1}{\sigma-1}}(1-\alpha)m_i^{-\frac{1}{\sigma}} = p_m$$

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. Moreover, recall that

$$m_i^{\frac{1}{\sigma}} = \left(\frac{z_i}{z_{i=0}}\right)^{\frac{\sigma-1}{\sigma}} m_{i=0}^{\frac{1}{\sigma}} \Rightarrow \tau_i = z_i^{\frac{\sigma-1}{\sigma}} p_m m_{i=0}^{\frac{1}{\sigma}}$$

[16]This setup is extendable to any arbitrary number of buyer layers.

Therefore, τ_i is distributed according to a Pareto distribution with tail parameter $\rho := \frac{\kappa\sigma}{\sigma-1}$. Notice that $\rho > \sigma$, since $\kappa > \sigma - 1$. Moreover, $h(\tau)$ is the hazard function of the τ distribution, given by:

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

Let D_m denote the total demand for intermediates in the economy summing both firm layers and all types within firm layers:

$$D_m = \sum_{i \in R,U} \int_{u_0} \left(\int_{z_{i0}} m_{iu}(z_i), \mu_i(z_i) N_i, dz_i \right) \mu_u N_u, du,$$

Denote

$$s_y = \frac{p_{u_0}\, y_{u_0}}{D_m}$$

as the revenue share of upstream firms with the smallest upstream type u_0.

Then, we have the optimal nonlinear price is characterized by the following solution:

$$T_{iu} = p_u m_{iu} + F_{iu}, \quad \text{with} p_u = \frac{\rho}{\rho - 1} c_u, \quad \rho = \frac{\sigma\kappa}{\sigma - 1},$$

The flat fee is given by:

$$F_{iu} = \text{Rev}_{i_0} \frac{1}{\left(\frac{\alpha}{1-\alpha}\right)^{\sigma} \left(\frac{p_m}{w}\right)^{\sigma-1} + 1} \left(\frac{1}{\sigma}\right) \left(\frac{p_u}{p_m}\right)^{1-\sigma} \qquad i \in \{r, u\}.$$

The revenue function is given by:

$$\text{Rev}_i = A_i q_i^{\theta}, \quad i \in \{r, u\},$$

where

$$\theta = \frac{\sigma - 1}{\sigma}, \quad A_r = Y^{\frac{1}{\sigma}}, \quad A_\zeta = \frac{D_m^{\frac{1-\sigma}{\sigma}} \left(p_m D_m + F_z N_r\right)}{1 - s_y \frac{1}{\left(\frac{\alpha}{1-\alpha}\right)^{\sigma} \left(\frac{p_m}{w}\right)^{\sigma-1} + 1}}.$$

Where F_u is the average flat fee received by upstream firm u from retailers.

The input demand functions satisfy:

$$m_{iu'} = \left(\frac{p_{u'}}{p_m}\right)^{-\sigma} m_i, \quad i \in \{r, u\},$$
$$m_i = \left(\frac{\alpha}{1-\alpha}\frac{p_m}{w}\right)^{-\sigma} l(\gamma), \quad i \in \{r, u\}.$$

Notice that it is always optimal for the seller firm to serve all buyer types. While excluding the lowest type $i_0, \quad i = r, u$ would allow the seller to charge a higher flat fee, doing so would result in the loss of demand associated with i_0. Under the Pareto distribution, the mass of firms of type i_0 is large enough that the seller never finds it optimal to exclude them. 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. 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 as the optiminal linear price proof basic construct.

Documentos de Trabajo Banco Central de Chile	Working Papers Central Bank of Chile
NÚMEROS ANTERIORES	PAST ISSUES
La serie de Documentos de Trabajo en versión PDF puede obtenerse gratis en la dirección electrónica: www.bcentral.cl/esp/estpub/estudios/dtbc. Existe la posibilidad de solicitar una copia impresa con un costo de Ch$500 si es dentro de Chile y US$12 si es fuera de Chile. Las solicitudes se pueden hacer por fax: +56 2 26702231 o a través del correo electrónico: bcch@bcentral.cl.	Working Papers in PDF format can be downloaded free of charge from: www.bcentral.cl/eng/stdpub/studies/workingpaper. Printed versions can be ordered individually for US$12 per copy (for order inside Chile the charge is Ch$500.) Orders can be placed by fax: +56 2 26702231 or by email: bcch@bcentral.cl.

DTBC – 1049
The Aggregate Welfare Effects of Nonlinear Prices in Supply Chains
Luca Lorenzini, Antonio Martner

DTBC – 1026*
The Incidence of Distortions
David Atkin, Baptiste Bernadac, Dave Donaldson, Tishara Garg, Federico Huneeus

DTBC – 1006*
Input Price Dispersion Across Buyers: New Evidence and Implications for Aggregate Productivity
Ariel Burstein, Javier Cravino, Marco Rojas

DTBC – 1048
The effect of automation on the labor market: An approach using firm-level microdata
Camilo Levenier

DTBC – 1047
Firm-level CO2 Emissions and Production Networks: Evidence from Administrative Data in Chile
Pablo Acevedo, Elías Albagli, Gonzalo García-Trujillo, María Antonia Yung

DTBC – 1046
The Cross Border Effects of Bank Capital Regulation in General Equilibrium
Maximiliano San Millán

DTBC – 1045
The impact of financial crises on industrial growth in the Middle East and North Africa
Carlos Madeira

DTBC – 1044
The impact of financial crises on industrial growth: lessons from the last 40 years
Carlos Madeira

DTBC – 1043
Heterogeneous UIPDs Across Firms: Spillovers from U.S. Monetary Policy Shocks
Miguel Acosta-Henao, Maria Alejandra Amado, Montserrat Martí, David Perez-Reyna

DTBC – 1042
Bank Competition and Investment Costs across Space
Olivia Bordeu, Gustavo González, Marcos Sorá

DTBC – 982*
Freight Costs and Substitution Among Import Regions: Implications for Domestic Prices
Gustavo González, Emiliano Luttini, Marco Rojas

DTBC – 1041
Tail-Risk Indicators with Time-Variant Volatility Models: the case of the Chilean Peso
Rodrigo Alfaro, Catalina Estefó

DTBC – 971*
Production Network Formation, Trade, and Welfare
Costas Arkolakis, Federico Huneeus, Yuhei Miyauchi

DTBC – 1011*
Macro Implications of Inequality-driven Political Polarization
Alvaro Aguirre

DTBC – 967*
Firm Financing During Sudden Stops: Can Governments Substitute Markets?
Miguel Acosta-Henao, Andrés Fernández, Patricia Gomez-Gonzalez, Sebnem Kalemli-Ozcan

DOCUMENTOS DE TRABAJO Julio 2025