Growing by the Masses
Revisiting the Link between Firm Size and Market Power*

Hassan Afrouzi†
Columbia University

Andres Drenik‡
Columbia University

Ryan Kim§
Johns Hopkins University

First Draft: August, 2020
This Draft: December, 2020

Abstract

How are firms' size and market power related? Using merged micro-data about producers and consumers, we show that a firm's size is mainly related to its number of customers, while its market power is associated only with average sales per customer. We study the macroeconomic implications of these facts by developing a firm dynamics model with customer acquisition and endogenous markups. By allowing firms to grow through customer acquisition, our model associates *higher* concentration at the top of the productivity distribution with a *lower* aggregate markup. Nonetheless, our quantitative analysis reveals (1) higher markup dispersion relative to conventional models, and (2) large welfare and efficiency losses due to misallocation of customers across firms. By concentrating customers among more productive firms, the efficient allocation achieves 10.8% higher aggregate productivity, 14.6% higher output and 13.6% higher welfare than the equilibrium allocation.

JEL classification: D61, D24, D43, E22.

Key Words: Misallocation, Customer acquisition, Markups, Concentration.

*We thank Costas Arkolakis, John Asker, David Baqaee, Gideon Bornstein, Ariel Burstein, Chris Edmond, Matthieu Gomez, Emilien Gouin-Bonenfant, Jennifer La'O, Virgiliu Midrigan, Chris Moser, Michael Peters, Tommaso Porzio, Jesse Schreger, Michael Woodford, and participants at the VMACs Junior Conference, as well as Cleveland Fed, Columbia Macro Lunch, Yale University, Johns Hopkins University, and University of Texas at Austin seminars for valuable comments and suggestions. We also thank Luigi Caloi for providing superb research assistance. Researchers own analyses calculated (or derived) based in part on data from The Nielsen Company (US), LLC and marketing databases provided through the Nielsen Datasets at the Kilts Center for Marketing Data Center at The University of Chicago Booth School of Business. The conclusions drawn from the Nielsen data are those of the researcher(s) and do not reflect the views of Nielsen. Nielsen is not responsible for, had no role in, and was not involved, in analyzing and preparing the results reported herein.

†hassan.afrouzi@columbia.edu.

‡ad3376@columbia.edu. 420 West 118th Street, New York, NY, 10027, USA.

§rkim59@jhu.edu.

1 Introduction

Canonical macroeconomic models of endogenous markups generate a close relationship between relative firm size and market power at the micro level (e.g., Atkeson and Burstein, 2008; Edmond, Midrigan and Xu, 2018), and a similarly tight link between concentration and misallocation at the macro level. While these relationships follow from the assumption that all firms face the same individual demand curve (henceforth, the intensive margin), in reality, firms spend vast resources to expand their customer bases, i.e., to shift their demand curves (henceforth, the extensive margin).[1] How does endogenous customer acquisition change our understanding of the relationship between concentration and misallocation?

In this paper, we answer this question by empirically and theoretically analyzing how each margin of demand contributes to firms' size and market power. First, we exploit a novel merged dataset between individual product-level consumption data and producer-level data to uncover the role of the intensive and extensive margins of demand for concentration and market power. Our motivational evidence reveals that while firms grow mostly through the extensive margin of demand, their market power is mainly associated with the intensive margin of demand. Second, we develop a theoretical framework with endogenous customer acquisition and variable markups that is consistent with these facts to study their implications for the misallocation of customers across firms, concentration, and market power. Third, we quantify efficiency losses from the misallocation of demand across firms and find that, relative to the efficient allocation, equilibrium aggregate TFP and output are 10.8% and 14% lower, respectively.

Our first contribution is to document the following three facts by merging the Nielsen Homescan Panel and Compustat datasets. First, differences in the size of firms' customer bases account for three-quarters of the variation in firms' sales. Second, firms' markups are not correlated with their number of customers. Instead, they are only positively associated with average sales per customer. Third, in assessing firms' abilities to shift their demand, we find that firms' non-production expenses are positively associated with the acquisition of new customers but not with the retention of their existing customers, nor with their average sales per customer. These motivational facts serve as the foundation of the theory that we develop to understand the relationship between different sources of firm size and market power.

In our model of firm dynamics, monopolistically competitive firms with heterogeneous productivity spend resources to acquire new customers in the presence of business-stealing externalities. Moreover, firms face a semi-kinked demand schedule from each customer, which implies a higher demand elasticity and lower market power at higher relative prices

[1] Arkolakis (2010) reports total spending in marketing of as high as 5% of GDP in the US.

(Kimball, 1995). Hence, while firms hold market power over each customer, the total number of customers only shifts their demand, as in Phelps and Winter (1970). These mechanisms link the model to our three motivating facts: our model allows firms to grow through both margins of demand (first fact), creates a comovement between markups and sales per customer, but not necessarily between markups and the number of customers (second fact), and implies an endogenous relationship between sales and non-production costs (third fact).

By allowing firms to grow through the extensive margin of demand, our model breaks the *direct* relationship between firm size and market power that is generated by the intensive margin of demand in canonical models. Since the extensive margin only shifts the firms' demand curves, it does not directly affect firms' market power. However, it allows for a relationship between size and market power through the costs and benefits of customer acquisition, as opposed to a direct relationship between demand elasticity and size. High-markup firms anticipate higher gains from additional customers, and invest more in their customer bases, subject to technological constraints on customer acquisition. Hence, once matched to the data, our model generates a similar relationship between firm size and market power, but one that is inherently different from conventional models and has notably different macroeconomic implications.

To study these macroeconomic implications, we characterize the efficient allocation in our model. Under the efficient allocation, the social planner increases aggregate productivity by allocating more customers towards more productive firms while equalizing the relative demand per customer across weakly substitutable varieties. This result contrasts with the efficient allocation in conventional models, where the planner has one instrument to target two mutually exclusive objectives: concentrate demand among more productive firms to increase aggregate productivity or equalize demand across varieties to increase utility from consumption. In our model, this trade-off is non-existent because the planner uses both margins of demand as instruments to achieve both objectives.

Even though our model implies lower distortions due to size differences across firms when they mainly grow through the extensive margin, welfare losses are potentially larger. This result follows from the observation that the endogenous allocation of customers pushes the Pareto frontier of our economy beyond what the conventional models would suggest. While the uniform allocation of customers across firms is still feasible, the planner improves on this allocation by concentrating customers among more productive firms. Hence, welfare losses can be large if the equilibrium allocation of customers is sufficiently distorted. Thus, our analysis unveils a novel source of efficiency losses due to the *misallocation of customers*.

Next, we calibrate the model to study its quantitative implications, and to measure the differences in the allocation of customers, aggregate productivity, and welfare between the

equilibrium and the efficient allocation. One of the key challenges in this analysis is to identify model parameters that determine the equilibrium allocation of customers across firms. To do so, we devise a strategy based on the model's predictions that we implement with available data on firms' sales and cost structures. At the core of this strategy is the comovement between a firm's sales and its non-production expenses (conditional on production expenses that control for confounding factors), which is informative of returns to scale in the customer acquisition technology.

With the calibrated model at hand, we conduct two exercises to analyze the micro- and macroeconomic consequences of endogenous customer acquisition in a model with variable markups. First, we provide comparative statics by comparing our equilibrium with those obtained in a restricted version of our model, in which customers are uniformly distributed across firms (as in conventional models). This restriction has large aggregate consequences. By forcing firms to switch their sales growth strategy from expanding their customer base to increasing their sales per customer, the top 5% sales share *declines* from 50% to 17%, but the aggregate markup *increases* by 12 percentage points. By giving more customers to less productive firms, relative to our baseline model, the restricted model features higher entry but much lower aggregate TFP and output. Thus, our model cautions against using measures of concentration to infer the degree of market power, and more generally, the degree of misallocation in the economy. Moreover, calibrating both models to the same set of moments, our model generates a higher degree of markup dispersion, which anticipates the large degree of misallocation that we find.

Our second quantitative result is that the misallocation of demand has large negative effects on efficiency and welfare: the consumption equivalent welfare gains of the representative household under the efficient allocation is 13.6%. The majority of this gain is coming from the efficiency gains in aggregate TFP under the planner's allocation, quantified at 10.8% higher than in the equilibrium. The planner achieves higher aggregate TFP by reallocating customers from low productivity firms to the most productive ones. Indeed, in the efficient allocation, the top 5% sales share increases by almost 40%, and the number of operating firms declines by 11%. Finally, we verify that these results are mainly driven by customer misallocation, which in the equilibrium is determined by the degree of decreasing returns to advertising. To do so, we show that by moving half-way from the calibrated model to an economy with constant returns to advertising, differences become much less pronounced. Compared to this new equilibrium allocation, the efficient allocation generates *only* 3.2% higher TFP, 4% higher welfare, and 15% higher concentration.

Literature review Our paper is closely related to canonical macroeconomic models of endogenous market power, such as Rotemberg and Woodford (1992), Atkeson and Burstein

(2008), and Edmond, Midrigan and Xu (2018), which predict a positive relationship between a firm's market share and its ability to exert market power.[2] Our contribution to this literature is twofold. From an empirical perspective, we provide evidence that market power is associated only with average sales per customer. From a theoretical perspective, we provide a theory that links market power to the empirically relevant measure of concentration. By endogenizing the extensive margin of demand in our framework, our calibrated model associates market concentration with higher markup dispersion but lower aggregate market power, relative to canonical models.[3]

Our model also relates to a large body of work that analyzes the consequences of endogenous customer acquisition through marketing or advertising activities (see, e.g., Arkolakis, 2010; Perla, 2019).[4] Moreover, Fitzgerald, Haller and Yedid-Levi (2016) and Fitzgerald and Priolo (2018) present empirical and model-based evidence in different settings in favor of these models. In a related theme, Kaplan and Zoch (2020) analyze the implications of expansionary activities of firms for the distribution of labor income. Finally, recent analysis by Einav, Klenow, Levin and Murciano-Goroff (2020) provide broader evidence of the importance of the size of customer base for firm size, and study the implications of customer acquisition in a growth model with constant markups. Relative to this literature, our main empirical contribution is to document the relationship between markups and different margins of demand. Furthermore, our main theoretical contribution is to analyze a model with variable markups and endogenous customer acquisition based on advertising activities. We embed this framework in a firm dynamics model (as in Hopenhayn, 1992) to study the misallocation of demand across firms by providing a comparison between the equilibrium and the efficient allocations.

Our paper is also related to the literature that analyzes the role of misallocation of production inputs across firms in affecting aggregate TFP (see the seminal work by Restuccia and Rogerson, 2008; Hsieh and Klenow, 2009). This literature has focused on multiple

[2] A similar relationship holds in models used in the international trade literature (see, e.g., Gopinath and Itskhoki, 2010; Hottman, Redding and Weinstein, 2016; Amiti, Itskhoki and Konings, 2019). See also Burstein, Carvalho and Grassi (2020) for novel evidence on how markups and relative size are related at the firm, industry and aggregate level.

[3] Thus, our model echoes a familiar argument against using measures of concentration to make predictions about changes in market power. A similar point has been raised by Syverson (2019), Neiman and Vavra (2019) and Covarrubias, Gutiérrez and Philippon (2020). For empirical analysis of the relationship between concentration and market power, see De Loecker, Eeckhout and Unger (2020), Crouzet and Eberly (2019) and Autor, Dorn, Katz, Patterson and Van Reenen (2020).

[4] See, also, Drozd and Nosal (2012); Sedláček and Sterk (2017). Furthermore, for models of customer acquisition through dynamic pricing see Phelps and Winter (1970); Bils (1989); Rotemberg and Woodford (1999); Ravn, Schmitt-Grohé and Uribe (2006); Nakamura and Steinsson (2011); Dinlersoz and Yorukoglu (2012); Gourio and Rudanko (2014); Foster, Haltiwanger and Syverson (2015); Cabral (2016); Gilchrist, Schoenle, Sim and Zakrajšek (2017); Hong (2017); Paciello, Pozzi and Trachter (2018); Bornstein (2018). We further discuss the relationship between our paper and these two branches of the literature in Section 3.7.

sources of misallocation: endogenous market power (Edmond et al., 2018; Peters, 2020), information frictions (David, Hopenhayn and Venkateswaran, 2016), adjustment costs (Asker, Collard-Wexler and De Loecker, 2014), financial frictions (Buera, Kaboski and Shin, 2011; Midrigan and Xu, 2014), and the interaction between frictions and the input-output structure of production (Baqaee and Farhi, 2019; Bigio and La'O, 2020). We contribute to this literature by highlighting a new source of distortions in aggregate productivity that stems from the misallocation of customers across firms.

Layout The paper is organized as follows. Section 2 describes the data and conducts the empirical analysis. Section 3 presents the model, and characterizes both the equilibrium and the efficient allocation. Section 4 presents the calibration strategy and results. Section 5 compares the baseline model with a restricted model to illustrate the role of endogenous customer acquisition. Section 6 quantifies efficiency losses. Finally, Section 7 concludes.

2 Motivating Facts

In this section, using micro-level data, we document three new facts that shed light on the importance of customer bases for firm dynamics and price-cost markups. To summarize, we find that:

Fact 1. Firms mainly grow by acquiring new customers, as opposed to increasing their average sales per customer.

Fact 2. Price-cost markups are correlated only with average sales per customer and are unrelated to the number of customers.

Fact 3. The acquisition of new customers is associated with the firm's non-production expenses, whereas average sales per customer are not.

Our theoretical analysis in Section 3 integrates these facts to understand the implications of customer acquisition for misallocation, market concentration and welfare.

2.1 Data Description

We construct a detailed customer-firm-matched dataset to decompose firms' sales into the size of their customer bases and average sales per customer. Formally, we consider the following exact decomposition of log sales of firm i ($\ln S_i$):

$$\ln S_i = \ln m_i + \ln (p_i q_i) , \tag{2.1}$$

where m_i is the number of customers firm i is facing, p_i is the price it charges, and q_i is the average quantity purchased per customer.

To measure the number of customers each firm has, we use the Nielsen Homescan Panel, made available by the Kilts Marketing Data Center at the University of Chicago Booth School of Business. The data contain approximately 4.5 million barcode-level product sales recorded from an average of 55,000 households per year in the United States. Nielsen samples households and provides in-home scanners to make those sampled households record their purchases of products with barcodes. A barcode is a unique universal product code (UPC) allocated to each product and is used to scan and store product information. Each household is assigned a sample weight—or a projection factor—by Nielsen based on ten demographic variables to make the sample nationally representative.[5] Nielsen assigns a broad product-group label for each product, such as pet food and school supplies, and records information about the retailer a household visited to purchase products at a given time. According to Nielsen, the Homescan Panel covers approximately 30 percent of all household expenditures on goods in the consumer price index (CPI) basket. The data we use covers the 2004-2016 period.

Next, we combine the Nielsen database with GS1 US Data Hub to group individual products according to their producing firms and merge in other firm-level information. GS1 is the business entity that provides barcodes to products and records the firm name for each UPC available in the Nielsen data. This allows us to link customer- and producer-level data for each product. The definition of a firm is based on the unit that purchased barcodes from GS1. Therefore, a firm in our data corresponds to either a manufacturer or a retailer.

Finally, we incorporate firm-level balance sheet information from Compustat to analyze firms' cost structures.[6] Compustat includes panel data on publicly traded firms since 1960. With the caveat that this dataset covers only publicly listed firms, it constitutes the main source of data for firm-level analysis in the US and has been used in the recent literature on price-cost markups (see, for example, De Loecker et al., 2020; Edmond et al., 2018; Traina, 2019). Throughout the analysis, we focus on two measures of a firm's costs from Compustat. From an accounting perspective, a firm's costs associated with the running of the firm are captured in the Operating Expense (OPEX), which is divided into the Cost of Goods Sold (COGS, production costs) and Selling, General, and Administrative Expenses

[5]The ten demographic variables are: household size, household income, head of household age, race, Hispanic origin, male head education, female head education, head of household occupation, the presence of children, and Nielsen county size.

[6]We match the Nielsen-GS1 database with the Compustat database, similar to what has been done in Argente, Lee and Moreira (2018). We use the "reclink" STATA software command based on company name after standardizing it with the "std_compname" command (Wasi and Flaaen 2015). Once Stata reports the matching rate for each observation, we keep those having higher than .99 matching rate. We manually check the company name for every observation and drop inconsistent matches.

(SGA, non-production costs). According to Compustat, COGS includes *"all expenses that are directly related to the cost of merchandise purchased or the cost of goods manufactured that are withdrawn from finished goods inventory and sold to customers"*. It records costs attributable to the production of the goods sold by a firm, and its typical categories are the cost of labor and intermediate inputs used in production. On the other hand, SGA expenses include *"all commercial expenses of operation (such as expenses not directly related to product production) incurred in the regular course of business..."*. It includes the costs incurred to sell and deliver products and services and the costs to manage the company, and typical categories are advertising, marketing, shipping, research and development, among others.

Table 1: Summary Statistics

Variable	**N**	**Mean**	**SD**	**p10**	**p50**	**p90**
	Panel A: Nielsen-GS1, Firm-Product Group-Year Variables					
S_{igt}	557,820	6,708.16	64,961.12	3.97	126.08	5,170.55
$p_{igt}q_{igt}$	557,820	10.03	20.14	1.96	5.88	19.50
m_{igt}	557,820	500.79	2,789.82	0.82	19.83	639.87
m_{igt}^{New}	557,820	250.34	988.95	0.48	16.03	424.89
m_{igt}^{Old}	557,820	250.45	1,963.09	0.00	1.60	194.18
	Panel B: Nielsen-Compustat, Firm-Year Variables					
SGA_{it}	2,101	2,009.17	4,993.82	7.63	299.09	4,882.24
COGS_{it}	2,299	7,147.11	18,558.75	17.17	1,123.66	17,251.26
OPEX_{it}	2,299	9,147.10	21,337.48	25.72	1,620.66	24,212.49
SGA-to-OPEX_{it}	2,101	0.30	0.20	0.08	0.27	0.58
$\text{Sales-to-COGS}_{it}$	2,299	1.79	1.06	1.14	1.49	2.61

Notes: The Nielsen-GS1 data in Panel A has 40,418 firms and 109 product groups in the period of 2005-2016. S_{igt} denotes sales of firm i in product group g and time t, $p_{igt}q_{igt}$ sales per customer, m_{igt} number of customers, m_{igt}^{New} new customers in year t who did not purchase products in year $t-1$, and m_{igt}^{Old} customers who purchase the products consecutively in year $t-1$ and t ($m_{igt} = m_{igt}^{\text{New}} + m_{igt}^{\text{Old}}$). S_{igt} is measured in thousands US dollar, and m_{igt}, m_{igt}^{New}, and m_{igt}^{Old} are in thousands of customers. All Nielsen variables are projection-factor adjusted. The Nielsen-Compustat data in Panel B has 332 firms in the period of 2005-2016. All cost-side variables are in millions US dollar and are deflated with the GDP deflator.

Table 1 presents summary statistics of the customer-firm matched data. Nielsen-GS1 data reveals that much of the firm-group-year sales is driven by the number of customers, not by average sales per customer: more than 500,000 customers spend only $10 approximately for each product group and firm per year on average. Among the total number of customers firms serve in product group g, approximately half of them are new customers

in any given year. Inspecting the SGA-to-OPEX ratio shows that non-production costs represent on average approximately 30% of the total costs and are an important component of a firm's total operating costs. However, there is a large degree of heterogeneity in firms' cost structure in the Nielsen-Compustat data. For example, the p90-p10 gap of the SGA-to-OPEX ratio is around 50%. Although we only have 332 firms in the Nielsen-Compustat matched data, these cover close to one-fourth of total sales in Nielsen. Appendix A provides a description of the data-cleaning procedure along with a more detailed description of the data.

2.2 Results

2.2.1 The Role of Customer Base for Sales Growth

We start by documenting that the main source of variation in firm sales is the variation in the number of customers, rather than average sales per customer. Following Equation (2.1), Table 2 decomposes the variance of log sales into the variances of log sales per customer and log number of customers, as well as the covariance between these two components. The number of customers accounts for approximately 80% of the variation in sales across firms. Average sales per customer accounts for approximately 11% of the variance of sales, and the covariance accounts for the rest.[7] These results parallel the findings in contemporary work by Einav et al. (2020). Using transaction-level data for a broad set of industries, they document that differences in customer bases account for 74% of sales variation across merchants. This shows that our finding extends beyond the consumer packaged goods sector and is representative of similar patterns in a wider set of industries.

$\text{Var}(\ln S_{igt})$	$\text{Var}(\ln p_{igt}q_{igt})$	$\text{Var}(\ln m_{igt})$	$2\text{Cov}(\ln p_{igt}q_{igt}, \ln m_{igt})$
7.5807	0.8672	6.1146	0.5989

Table 2: Decomposing the Variance of Sales

Notes: S_{igt} denotes sales, $p_{igt}q_{igt}$ sales per customers, and m_{igt} the number of customers. We use 557,820 firm-group-year-level observations in Nielsen-GS1 data. Sales and the number of customers are projection-factor adjusted.

In addition to decomposing sales in the cross-section of firms, we find that the acquisition of new customers is also the main driver of firms' sales growth. As firms enter the economy, they can grow in two ways: either by selling more per customer or by selling to more customers. To document this fact, we analyze firm growth patterns after entry. We

[7] Our decomposition results are similar when we instead use the first-difference of log sales; approximately 78%, 20%, and 2% of the variation are explained by the Δ log number of customers, Δ log average sales per customer, and the covariance between the two, respectively.

mark a firm's entry as the time when it appears in our data for the first time. To be conservative, we drop entry events that occurred in the first four years in the dataset.[8] To quantify the importance of each margin, we estimate the following equation:

$$\ln S_{it} = \sum_{a=1}^{8} \delta_a \mathbf{1}(\text{age}_{it} = a) + \lambda_i + \lambda_t + \varepsilon_{it},$$

where S_{it} stands for sales and its components of firm i in year t, age_{it} is the number of years firm i stayed in the economy after entry in year t, and λ_i and λ_t are the firm and year fixed effects, respectively (see Argente, Lee and Moreira (2019) for a similar analysis of the life-cycle of individual products). The parameters of interest are δ_a, which measures the dynamics of average sales and its components over the life-cycle of the firm.

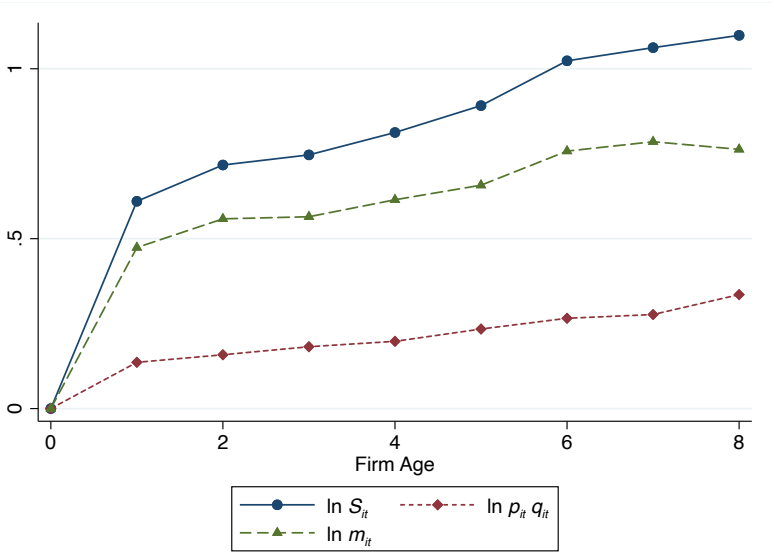


Figure 1: Decomposition of Firm Sales Growth by Firm Age

Notes: This figure plots the average firm sales, sales per customers, and number of customers for each firm-age based on Equation (2.2.1), after controlling for firm and year fixed effects. The blue circled line denotes average log sales. The red diamond line denotes average log sales per customer. The green triangle line denotes the average log number of customers as a dependent variable. There are 40,442 observations and 9,990 firms that newly enter the economy starting from the year 2008 in the Nielsen-GS1 data. All estimates are normalized based on age 0. All variables are projection-factor adjusted.

Regardless of the firm age, sales growth is mostly attributed to the increase in the number of customers. Figure 1 plots the average log sales as a function of firm age (δ_a) and decomposes it into the average log number of customers and the average log sales per cus-

[8]There is a large increase in the number of households and firms in the Nielsen Homescan Panel data in the years 2006 and 2007. We choose the years 2004-2007 as the initial period to make the analysis conservative. Thus the maximum firm age in our sample of entrants is eight years.

tomer. At age 1, differences in the number of customers explain approximately 78% of differences in sales, whereas sales per customer explain approximately 22% of sales. Although the importance of the number of customers decreases as firms become older, on average, it still explains approximately 70% of sales for the maximum firm age observed in the data. Results are robust to including only those firms that survive at least 3 or 5 consecutive years and analyzing average monthly sales as the dependent variable, which accounts for entry throughout the year. Since the degree of durability of a product might affect the ability of firms to grow through different margins, we repeat the analysis by splitting products according to their durability and find similar patterns within both subsamples. See Appendix B for further details.

2.2.2 The Relationship between Firm Size and Markups

Armed with the empirical evidence showing that differences in firms' size stem mainly from differences in the size of their customer bases, we revisit the predictions of a large class of models that relate a firm's size with its market power (see, e.g., Rotemberg and Woodford, 1992; Atkeson and Burstein, 2008; Edmond et al., 2018). These models predict that larger firms charge higher markups. Our decomposition of firms' sales in Equation (2.1) raises the following question: which margin of sales captures the relationship between size and markups? We answer this question by estimating the following regression equation:

$$\ln \text{Markup}_{it} = \alpha_1 \ln p_{it} q_{it} + \alpha_2 \ln m_{it} + \lambda_{s,t} + \varepsilon_{it}, \tag{2.2}$$

where Markup_{it} is the price-cost markup charged by firm i at time t. The sector-year fixed effects $\lambda_{s,t}$ absorb all the variation at the sector-year-level, which allows us to interpret markups and measures of size in *relative* terms. To measure markups we follow the methodology by De Loecker et al. (2020), where markups are equal to the inverse variable cost share of sales multiplied by the output elasticity with respect to those variable inputs. Because this methodology does not require information on all variable costs, we follow De Loecker et al. (2020) and use data on COGS from Compustat as a measure of variable costs. In addition, since we are interested in relative markups within industries at a given point in time, our specification in Equation (2.2) absorbs the output elasticity with the set of fixed effects.[9]

We find that firms' markups are highly correlated with their sales per customer but are

[9]The main challenge in the estimation of markups lies in the estimation of the output elasticity using only data on firms' revenues (see Bond, Hashemi, Kaplan and Zoch, 2020). Given our focus on relative markups, we do not need to estimate output elasticities. Instead, our regression specification incorporates a set of fixed effects which absorb these output elasticities. The underlying assumption we make, which is standard in the literature, is that the output elasticity with respect to COGS remains constant across firms within an industry and/or time.

	ln Markup$_{it}$				
	(1)	(2)	(3)	(4)	(5)
ln $p_{it}q_{it}$	0.092***	0.091***	0.060***	0.059***	0.060**
	(0.033)	(0.033)	(0.022)	(0.022)	(0.024)
ln m_{it}	-0.002	-0.002	0.002	0.002	0.003
	(0.006)	(0.006)	(0.007)	(0.007)	(0.007)
Observations	2433	2433	2433	2433	2433
R^2	0.046	0.047	0.311	0.313	0.338
Year FE		✓		✓	
SIC FE			✓	✓	
SIC-year FE					✓

Table 3: Markups, Sales per Customer, and Number of Customers

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$; standard errors are clustered at the firm-level. Markups are measured as the Sales-to-COGS ratio. The variable ln $p_{it}q_{it}$ denotes the log of the average sales per customers and ln m_{it} denotes the log number of customers. SIC industries corresponds to a two-digit SIC code. All Nielsen variables are projection-factor adjusted.

unrelated to their number of customers, as reported in Table 3. Results are robust to including different combinations of year and sector fixed effects. With the inclusion of industry-time fixed effects, our results show that firms that charge higher markups have a higher market share in terms of sales per customer, not in terms of the number of customers. That is, the relevant notion of market share is based on sales per customer. Appendix B presents results of an alternative specification that replaces the number of customers with a firm's sales. The greater importance of sales per customer in explaining markups remains.

Additional Evidence Based on Product-level Markups Our results in Table 3 are limited by the scope of the Compustat dataset, which focuses on large public firms and firm-level markups. To demonstrate the external validity of these results, we perform two complimentary analyses by using alternative measures of markups and samples. In the first approach, we measure the retailer-product-level markup as the difference between the retailer-product-level price available from the Nielsen Homescan Panel data and the wholesale cost obtained from the Nielsen PromoData, which is similar to the approach followed in Gopinath, Gourinchas, Hsieh and Li (2011) and Stroebel and Vavra (2019). In the second approach, we analyze the relationship between markups and each margin of demand by exploiting the rich dimensions of our data and controlling for marginal costs through an extensive set of fixed effects. This approach is similar to that of Fitzgerald et al. (2016) and is valid under the assumption of common marginal costs across different subsets of observations. Appendix B.1.3 provides detailed descriptions of both approaches.

Although we switch the focus to product-level markups and a broader sample of products and firms, both of our alternative approaches generate results consistent with those in Table 3: markups are positively associated with average sales per customer, but not with the number of customers.

2.2.3 Customer Acquisition and Firms' Non-production Costs

Our last set of results examine the scope of firms' control to grow through different margins of sales. For this, we use data on SGA expenses, which capture firms' expenses on expansionary activities, and provide empirical evidence indicating that: (1) firms that spend more non-production costs have larger sales, and (2) these costs are associated with the number of new customers firms acquire, but not with the number of old customers or average sales per customer. More specifically, we estimate the following specification:

$$\ln S_{igt} = \gamma \ln \text{SGA}_{it} + \boldsymbol{X}_{\boldsymbol{it}}'\boldsymbol{\nu} + \lambda_{ig} + \lambda_{st} + \lambda_{gt} + \varepsilon_{igt},$$

where S_{igt} stands for sales and its components of firm i in product group g and year t, $\boldsymbol{X}_{\boldsymbol{it}}'$ is a vector of firm-time-level control variables, λ_{ig} are firm-product-group fixed effects, λ_{st} are 2-digit SIC-year fixed effects, and λ_{gt} are product-group-year fixed effects. We allow for both product-group fixed effects and firm-SIC-code fixed effects to compare products within fine product categories. The vector of controls $\boldsymbol{X}_{\boldsymbol{it}}'$ includes lagged total sales and lagged total number of customers, which allow us to compare firms with similar sizes and customer bases. The coefficient of interest is γ, which captures the correlation between total sales (and its components) and SGA expenses.

As shown in the first column of Table 4, firms that spend more on SGA expenses have larger sales. Moreover, the second and third columns show that approximately 95% (0.090/0.095) of the correlation between sales and SGA expenses is due to the correlation between the non-production costs and the number of customers, not to the correlation with the average sales per customer. Finally, the last two columns further decompose the correlation of SGA expenses with the size of firms' customer bases into the acquisition of new customers and the retention of old customers. We find that, while there is a strong correlation between SGA expenses and the number of new customers, the regression coefficient for old customers is neither economically nor statistically significant. These results show that non-production costs of firms are associated with the acquisition of new customers, rather than maintaining the existing customer base.[10]

We have shown that firms' SGA expenses contribute to firm size only through customer

[10] As shown in Appendix B, we find similar results in our subsamples of durable and non-durable products.

	Decomposition of $\ln S_{igt}$			$\ln m_{igt}$: New vs. Old	
	(1) $\ln S$	(2) $\ln pq$	(3) $\ln m$	(4) $\ln m^{New}$	(5) $\ln m^{Old}$
$\ln SGA_{it}$	0.095***	0.005	0.090***	0.095***	0.016
	(0.036)	(0.014)	(0.028)	(0.032)	(0.027)
Observations	13131	13131	13131	13131	13131
R^2	0.962	0.909	0.965	0.943	0.961
Firm-year Controls	✓	✓	✓	✓	✓
Group-year FE	✓	✓	✓	✓	✓
SIC-year FE	✓	✓	✓	✓	✓
Group-firm FE	✓	✓	✓	✓	✓

Table 4: Sales and SGA: Decomposition

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$; standard errors are two-way clustered at the firm and product group level. S denotes total sales, pq the average sales per customers, m the number of customers, m^{New} the number of new customers, and m^{Old} the number of old customers. New customers are defined as the customers who do not purchase firm i's products in group g at time $t-1$ but start to purchase those products at time t, whereas old customers are the customers who consecutively purchase firm i's products in group g in both $t-1$ and t ($m = m^{New} + m^{Old}$). SIC industries corresponds to a two-digit SIC code. All Nielsen variables are projection-factor adjusted.

acquisition. Since firms' markups do not depend on the size of their customer base, the extent of firms' abilities to grow through investing in their customer bases weakens the relationship between firm size and market power. Anticipating our results in the theory, the extent to which firms can grow through customer acquisition depends on a *variable* component of SGA expenses, which has been a topic of discussion in the recent literature (see Traina, 2019; De Loecker et al., 2020). The correlation between sales and SGA expenses in Table 4 is indicative of such a variable nature of these costs. However, as we show in Appendix B, this contemporaneous correlation is weaker than the one between sales and COGS—which is commonly considered as a measure of variable production costs. Therefore, total SGA expenses seem to be composed of both variable and fixed components. While the Compustat dataset is not detailed enough to separate these components, in our model we incorporate both components and provide a strategy to measure how much firms can grow through investing in their customer bases.

3 Model

Time is discrete and is indexed by $t \in \{0, 1, 2, \dots\}$. The economy consists of a representative household with a unit measure of individual members denoted by $j \in [0, 1]$ and a measure of firms that operate in a representative industry and produce weakly substitutable goods.

We index firms by $i \in N_t$, where N_t is the set, and with a slight abuse of notation, the measure of producing firms at time t. We assume that only customers of a firm can purchase its product.

3.1 Households

At any given time, the representative household supplies labor to the firms in a competitive labor market and forms demand for the varieties produced by firms, taking their prices as given. We let $m_{i,t}$ denote both the measure and the set of variety i's customers, and write $j \in m_{i,t}$ when member j is a customer for variety i.

Preferences The household members jointly maximize their utility using a *Kimball aggregator* for aggregating their consumption utility. They solve:

$$\max_{\{C_t, L_t, (c_{i,j,t})_{j\in m_{i,t}}^{i\in N_t}\}} \sum_{t=0}^{\infty} \beta^t \left[\frac{C_t^{1-\gamma}}{1-\gamma} - \xi \frac{L_t^{1+\psi}}{1+\psi}\right] \tag{3.1}$$
$$s.t. \int_0^{N_t}\int_0^1 \mathbf{1}_{\{j\in m_{i,t}\}} \Upsilon\left(\frac{c_{i,j,t}}{C_t}\right) djdi = 1$$
$$\int_0^{N_t}\int_0^1 p_{i,t} c_{i,j,t} djdi \leq W_t L_t + \int_0^{N_t} \Pi_{i,t} di - T_t,$$

where $c_{i,j,t}$ is the consumption of member j from variety i, $p_{i,t}$ is the price of variety i, C_t is the household's aggregate consumption, L_t is total labor supply of the household, W_t is the wage, $\Pi_{i,t}$ is the profit of firm i and T_t is an aggregate lump-sum tax. Moreover, the function $\Upsilon(.)$ is strictly increasing and strictly concave with $\Upsilon(1) = 1$.[11]

Demand for Varieties The first property of the demand is that all the customers of variety i choose to purchase the same amount of it, where the other members of the household purchase none:

$$\frac{c_{i,j,t}}{C_t} = \begin{cases} \Upsilon'^{-1}\left(\frac{p_{i,t}}{P_t D_t}\right) & j \in m_{i,t} \\ 0 & j \notin m_{i,t} \end{cases}$$

Here, $D_t \equiv \left[\int_{i\in N_t}\int_0^1 \mathbf{1}_{\{j\in m_{i,t}\}} \frac{c_{i,j,t}}{C_t} \Upsilon'\left(\frac{c_{i,j,t}}{C_t}\right) djdi\right]^{-1}$ is an *aggregate demand index* and P_t is the price of the aggregate consumption good, which, henceforth, we normalize to one.[12] Therefore, the household's total demand for variety i is *proportional* to the number of its customers

[11] In the case of the CES aggregator, $\Upsilon(x) = x^{1-\sigma^{-1}}$, where σ is the elasticity of substitution across varieties.

[12] In the special case where the aggregator is CES, this demand index takes a value of $1/(1-\sigma^{-1})$; however, with the generalized Kimball aggregator this quantity is not necessarily a constant. Moreover, one could characterize the equations that pin down P_t and D_t in terms of prices rather than quantities. These equations

and is characterized by the following demand function:

$$c_{i,t} \equiv \int_0^1 \mathbf{1}_{\{j \in m_{i,t}\}} c_{i,j,t} dj = m_{i,t} q_{i,t} C_t \tag{3.2}$$

$$q_{i,t} \equiv \Upsilon'^{-1}\left(\frac{p_{i,t}}{P_t D_t}\right), \tag{3.3}$$

where $c_{i,t}$ is total demand for variety i, and $q_{i,t}$ is the *relative demand per customer* of variety i. The expression in Equation (3.2) maps our theory of demand to the starting point of our empirical analysis in Equation (2.1), where we decomposed the sales of a firm to its number of customers times its sales per customer. In our model, the same decomposition holds:

$$\ln(S_{i,t}) = \ln m_{i,t} + \ln p_{i,t} \Upsilon'\left(\frac{p_{i,t}}{D_t}\right) + \ln C_t.$$

It is important to note that this decomposition relies on the homogeneity of household members' preferences for the variety of firm i, or alternatively on the lack of sorting in how customers are matched to certain firms. In Appendix C, we introduce an extension of the model with heterogeneity in tastes as well as sorting and provide motivational evidence for why we abstract from them.

Elasticities and Super-elasticities of Demand To generate variable markups, following Edmond et al. (2018), we use the Kimball aggregator function from Klenow and Willis (2016) for our quantitative analysis:

$$\Upsilon(q) = 1 + (\sigma - 1) e^{\frac{1}{\eta}} \eta^{\frac{\sigma}{\eta} - 1} \left[\Gamma\left(\frac{\sigma}{\eta}, \frac{1}{\eta}\right) - \Gamma\left(\frac{\sigma}{\eta}, \frac{q^{\frac{\eta}{\sigma}}}{\eta}\right)\right],$$

where $\Gamma(.,.)$ is the incomplete Gamma function.[13] Intuitively, Kimball demand is a smoothed version of a kinked demand curve (Dotsey and King, 2005; Basu, 2005) where the relative demand is more elastic to the price at higher relative prices. Implementing this functional form in Equation (3.3), we can derive the following expression for the relative demand per

are:

$$\int_0^{N_t} m_{i,t} \Upsilon\left(\Upsilon'^{-1}\left(\frac{p_{i,t}}{P_t D_t}\right)\right) di = 1$$

$$\int_0^{N_t} m_{i,t} \frac{p_{i,t}}{P_t} \Upsilon'^{-1}\left(\frac{p_{i,t}}{P_t D_t}\right) di = 1,$$

which jointly determine P_t and D_t.

[13] $\Gamma(s,x) \equiv \int_x^\infty t^{s-1} e^{-t} dt$.

customer for firm i at time t:

$$q_{i,t} = \left[1 - \eta \ln\left(\frac{p_{i,t}}{D_t(1-\sigma^{-1})}\right)\right]^{\frac{\sigma}{\eta}}, \tag{3.4}$$

Moreover, this specification for $\Upsilon(.)$ is a generalization of the CES case with parameters $\sigma > 1$ and $\eta \geq 0$ that vary the *elasticity* and *super-elasticity* of demand. Formally, these quantities are given by

$$\varepsilon_{i,t} \equiv -\frac{\partial \ln(c_{i,t})}{\partial \ln(p_{i,t})} = \sigma q_{i,t}^{-\frac{\eta}{\sigma}}, \quad \varepsilon^{\varepsilon}_{i,t} \equiv -\frac{\partial \ln(\varepsilon(q_{i,t}))}{\partial \ln(p_{i,t})} = \eta q_{i,t}^{-\frac{\eta}{\sigma}},$$

where $\varepsilon_{i,t}$ is the expression for the *elasticity of demand* and shows that with the Kimball aggregator, the demand elasticity is a decreasing function of the relative demand per customer. In the special case when the super-elasticity of demand approaches zero ($\eta \to 0$), we are back to the standard case of the CES aggregator with σ being the constant elasticity of substitution across varieties.

Dynamics of Customer Bases Firms launch advertising campaigns to attract new customers from a pool of newly separated consumers. Two processes in the model cause separation and make potential customers available to all firms: at the end of each period, (1) all customers of exiting firms separate, and (2) customers of incumbent firms separate at an exogenous rate of $\delta \in [0, 1]$. Furthermore, we assume that the total number of matches that can be generated is fixed and exogenous to the advertising choices of firms. Without loss of generality, we normalize the total mass of matches to 1. This implies that while advertising affects the distribution of customers across firms, it does not increase the total number of customers that buy from an industry (see, e.g., Einav et al., 2020, for a similar assumption). We discuss this assumption further in Section 3.7.

Formally, we assume that operating firm i at time t posts $a_{i,t} \geq 0$ ads to acquire new customers. Every available member then draws one ad from the pool of all available ads and is matched to the firm that they draw. Therefore, the number of new customers that firm i acquires at time t is proportional to the number of ads that it posted, but it is normalized by the total number of ads posted by all firms as well as the pool of available members at any given time. Hence, firm i's customer base evolves according to:

$$m_{i,t} \leq (1-\delta)m_{i,t-1} + \frac{a_{i,t}}{P_{m,t}}, \tag{3.5}$$

where the inequality captures the notion that there is free disposal of customers should the firm choose to exercise that option. Moreover, we interpret $P_{m,t}$ as the *cost of a new customer*

in units of ads: it is the number of ads that a firm needs to post to get one new customer and is determined in the equilibrium so that the matching market clears ($\int_0^1 m_{i,t}di = 1$):

$$P_{m,t} = \frac{\int_0^{N_t} a_{i,t}di}{1-(1-\delta)\int_0^{N_t} m_{i,t-1}di}.$$

This expression shows that the cost of a match decreases with the total number of separated customers and increases with the total number of posted ads by all firms.

Labor Supply The household's labor supply is characterized by the following standard intra-temporal Euler equation: $\xi L_t^{\psi} = W_t C_t^{-\gamma}$.

3.2 Firms

On the firm side, we assume a structure with endogenous entry and exit, with the following timeline of events—as summarized in Figure 2. At the beginning of each period, a set of potential entrants are born with an initial productivity and decide whether to enter or not. Incumbents also draw a new productivity in the beginning of each period and decide whether to stay or exit. After entry and exit decisions are made, all firms pay an overhead cost of operation, decide on advertising campaigns, set prices and produce to meet demand. In the rest of this section, we provide a detailed description of these decisions.

Entry and Exit Decisions At each period t, a measure λ of potential entrants are born, each with an initial productivity $z_{i,t}$ drawn from a log-normal distribution:

$$\ln(z_{i,t}) \sim \mathcal{N}(\bar{z}_{ent}, \sigma_z^2). \tag{3.6}$$

We let Λ_t denote the set of these potential entrants at t. Incumbents, i.e., firms that entered the economy at least one period ago, also draw new productivities according to the following AR(1) process:

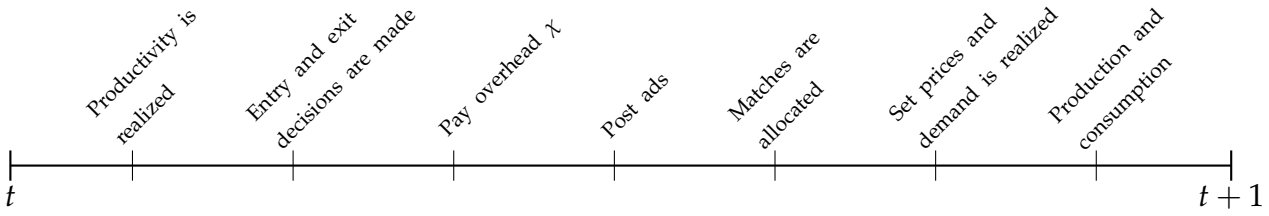
$$\ln(z_{i,t}) = \rho \ln(z_{i,t-1}) + \epsilon_{i,t}, \quad \epsilon_{i,t} \sim \mathcal{N}(0, \sigma_z^2) \tag{3.7}$$

With new productivities drawn, each incumbent or potential entrant then decides whether to stay in the economy or to dropout.[14] We refer to this decision by $\mathbf{1}_{i,t} \in \{0,1\}$, with 1

[14]Following Clementi and Palazzo (2016) and Ottonello and Winberry (2018), we allow for the mean of incumbents' productivity distribution—normalized to 0—to be different than that of entrants, $\bar{z}_{ent}$. This introduces a natural trend in firms' productivity based on their age and allows us to account for differences in size across age-groups, as reported in the Business Dynamics Statistics (BDS).

being an indicator for entering or staying. Finally, all the incumbent firms who decided to stay draw Bernoulli survival shocks, $\nu_{i,t}$, that are equal to 1 with probability $\nu \in [0,1]$, and drop out if $\nu_{i,t} = 0$. We assume $\nu_{i,t}$ is i.i.d. across firms and time.

Figure 2: Timing of Events



Notes: The figure shows the timing of firms' decisions in the model.

Advertising, Pricing and Production Decisions Firms who stay or enter the economy pay an overhead cost of $\chi > 0$ in units of labor at each period, which allows them to market and produce their product using labor. In particular, firms can use labor to produce ads using technology $a_{i,t} = l_{i,s,t}^{\phi} \geq 0$, where $l_{i,s,t}$ denotes the amount of labor allocated to advertising activities. The firm's customer base then evolves according to the law of motion in Equation (3.5), where $m_{i,t-1} \equiv 0$ for operating firms that entered at time t. Moreover, $\phi \in [0,1]$ is the degree of decreasing returns to advertising.

Furthermore, for a given number of customers, a firm's demand is given by Equation (3.2). Firms take this demand schedule as given and choose the price that maximizes their life-time profits. Each firm i then produces to meet its realized demand using technology $y_{i,t} = z_{i,t} l_{i,p,t}^{\alpha}$, where $z_{i,t}$ is the firm's productivity, and $l_{i,p,t}$ is its labor demand for production. Finally, $\alpha \in [0,1]$ is the degree of decreasing returns to production.

Firms' Problem Given an initial level of productivity and customer base, firm i's problem is given by

$$v_t(m_{i,t-1}, z_{i,t}) \tag{3.8}$$

$$\equiv \max_{(p_{i,\tau}, l_{i,s,\tau}, l_{i,p,\tau}, \mathbf{1}_{i,\tau})_{\tau=t}^{\infty}} \mathbb{E}_t \sum_{\tau=t}^{\infty} (\beta\nu)^{\tau-t} \left(\prod_{h=t}^{\tau} \mathbf{1}_{i,h} \right) \left(\frac{C_\tau}{C_t} \right)^{-\gamma} \left[\underbrace{p_{i,\tau} y_{i,\tau}}_{\text{total sales}} - \underbrace{W_\tau l_{i,p,\tau}}_{\text{COGS}} - \underbrace{W_\tau (l_{i,s,\tau} + \chi)}_{\text{SGA expenses}} \right]$$

$$\text{subject to} \quad y_{i,\tau} = m_{i,\tau} q_{i,\tau} C_\tau = z_{i,\tau} l_{i,p,\tau}^{\alpha} \tag{3.9}$$

$$q_{i,\tau} = \left[1 - \eta \ln \left(\frac{p_{i,\tau}}{D_\tau (1 - \sigma^{-1})} \right) \right]^{\frac{\sigma}{\eta}} \tag{3.10}$$

$$m_{i,\tau} \leq (1-\delta)m_{i,\tau-1} + \frac{l_{i,s,\tau}^{\phi}}{P_{m,\tau}}, \quad l_{i,s,t} \geq 0. \tag{3.11}$$

The problem states that firm i maximizes the expected discounted stream of its profits subject to Equation (3.9) that requires firm to meet its demand given its choice of a price, Equation (3.10) that specifies demand per customer under the Klenow and Willis (2016) specification for the Kimball aggregator, and Equation (3.11) that captures the law of motion for the number of firm's customers and the non-negativity of labor allocated towards advertising. Also, notice that we have separated the terms in the profit function to *total sales*, *COGS*, and *SGA expenses*, all of which we use later to discuss how we map the model to the data.

3.3 Characterization of Firms' Decisions

In this section, we characterize the firms' optimal decision rules for pricing, advertising, entry and exit.

Prices and Markups For a firm that has decided to operate in a given period, and for a given choice of advertising that determines its number of customers in that period, its pricing decision has a static nature. Formally, the firm chooses to charge an optimal markup over its marginal cost of production:

$$p_{i,t} = \underbrace{\frac{\varepsilon_{i,t}}{\varepsilon_{i,t}-1}}_{\text{markup}} \times \underbrace{\alpha^{-1}\frac{W_t l_{i,p,t}}{y_{i,t}}}_{\text{marginal cost}}. \tag{3.12}$$

This expression shows that despite the presence of variable advertising costs, the common proportionality relationship derived in conventional models between the labor share and the markup also holds in our model. This justifies our use of the De Loecker et al. (2020) methodology in identifying markups from the Compustat data.

It is also important to note that the firm's elasticity of demand, $\varepsilon_{i,t}$, is itself a function of demand per customer in Equation (3.4) and varies with the firm's pricing choice. Therefore, as long as η is not zero, the optimal markup of the firm varies with its marginal cost, which leads to the following lemma.

Lemma 1. *Firms with higher marginal costs charge higher prices and lower markups. Formally, let $\mu_{i,t}$ denote a firm's markup and $mc_{i,t}$ denote its marginal cost. Then, the elasticities of markups and*

prices to marginal costs are:

$$d\ln\left(\frac{p_{i,t}}{D_t}\right) = \frac{1}{1+\eta\sigma^{-1}\varepsilon_{i,t}(\mu_{i,t}-1)} d\ln\left(\frac{mc_{i,t}}{D_t}\right) \tag{3.13}$$

$$d\ln(\mu_{i,t}) = -\frac{\eta\sigma^{-1}\varepsilon_{i,t}(\mu_{i,t}-1)}{1+\eta\sigma^{-1}\varepsilon_{i,t}(\mu_{i,t}-1)} d\ln\left(\frac{mc_{i,t}}{D_t}\right) \tag{3.14}$$

Proof. See Appendix D.1. □

Equation (3.13), which is also known as the *incomplete pass-through* property of Kimball demand (see, e.g., Gopinath and Itskhoki, 2010; Amiti et al., 2019), shows that for a one percent increase in the marginal cost, the firm's relative price also increases but by less than one percent. The intuition for this result is that firms with higher marginal costs have to charge higher relative prices to make a positive margin, but since at higher prices demand is more elastic, their optimal markups are lower—in particular, this contrasts with a CES demand system, which requires the elasticity of demand to be constant all across the firms' demand curves.

We are now ready to prove the following proposition that links the model to our empirical fact in Table 3.

Proposition 1. *Firms with higher sales per customer charge higher markups. Formally,*

$$d\ln(\mu_{i,t}) = \eta\sigma^{-1}\mu_{i,t}(\mu_{i,t}-1) d\ln\left(\frac{p_{i,t}q_{i,t}}{D_t}\right) \tag{3.15}$$

Proof. See Appendix D.2. □

While the relationship outlined by Proposition 1 is independent of firms' customer base, the relationship between markups and total sales depends on how firms' customer bases covary with markups. Definite statements about this relationship require characterizing the optimal advertising strategies of firms; however, the results above are enough to make comparisons across firms with the same size and productivity:

Corollary 1. *Conditional on the same level of total sales and productivity, firms with a larger customer base charge lower markups—only because they face higher marginal costs.*

Proof. See Appendix D.3. □

Corollary 1 follows from the fact that once we fix productivity and total sales, firms with a larger number of customers should be selling less per customer. This highlights the main departure of our paper from the literature on variable markups, where customer acquisition is not modeled explicitly. These models often implicitly assume that customers

are homogeneously distributed across firms (i.e., a representative consumer buys from all firms). Hence, in those models, larger firms are larger because of their higher sales per customer, which creates an unbreakable link between markups and size with the Kimball aggregator or kinked demand curves. However, in our model, firms can be large either because they sell more per customer—hence, charging higher markups—or because of having a larger number of customers, which is associated with *lower* markups once we control for productivity and sales.

Advertising Strategies A key feature of our model is that firms internalize the decision of acquiring customers and can spend resources to do so. For this decision, while the marginal cost is determined by the amount of labor that the firm needs to utilize to find a new customer, its benefit is closely linked to the firm's market power and the amount of relative demand per customer. The following proposition formulates the optimality condition for firms' advertising decisions in terms of this cost-benefit analysis.

Proposition 2. *The optimal advertising strategy of a firm is characterized by*

$$\underbrace{\phi^{-1}\frac{W_t l_{i,s,t}}{m_{i,t}-(1-\delta)m_{i,t-1}}}_{\text{marginal cost of a new customer}} = \mathbb{E}_t \sum_{\tau=t}^{\infty} \underbrace{\left[(\nu(1-\delta))^{\tau-t}\prod_{h=t}^{\tau}\mathbf{1}_{i,\tau}\right]}_{\text{probability of match survival}} \underbrace{\beta^{\tau-t}\left(\frac{C_\tau}{C_t}\right)^{-\gamma}(\mu_{i,\tau}-1)\,mc_{i,\tau}q_{i,\tau}C_\tau}_{\text{discounted (gross) marginal profit per customer}}. \tag{3.16}$$

Proof. See Appendix D.4. □

Equation (3.16) shows that the marginal benefit of acquiring one more customer is linked to the net present value of the gross profits that the firm will earn from that customer for the duration of the match.

It follows from Proposition 1 that the marginal profits generated by a new customer are increasing in the markup of the firm. Therefore, firms that charge higher markups (or expect to charge higher markups on average for the duration of a match) anticipate a higher return on investing in their customer base. Hence, our model predicts a positive relationship between markups and the size of firms' customer bases. This can be more easily seen for the special case with $\delta = 1$, which allows for an analytical solution for the optimal customer base $m_{i,t}$ as a function of static profits per customer:

$$m_{i,t} = \frac{[(1-\mu_{i,t}^{-1})p_{i,t}q_{i,t}]^{\frac{1}{\phi^{-1}-1}}}{\int_{i\in N_t}[(1-\mu_{i,t}^{-1})p_{i,t}q_{i,t}]^{\frac{1}{\phi^{-1}-1}}di},$$

where the numerator is increasing in the firm's markup by Proposition 1.

Entry and Exit Policies A potential entrant enters the economy and an incumbent decides to stay if their value, specified in Equation (3.8) is positive: $v_t(m_{i,t-1}, z_{i,t}) \geq 0$. It can be shown that firms' value functions are increasing in their productivity and hence, for any given level of m_{-1}, there is a threshold productivity $z^*(m_{-1})$ such that firms with productivity higher than $z^*(m_{-1})$ stay or enter the economy (Hopenhayn, 1992).

3.4 Equilibrium

A monopolistically competitive equilibrium for this economy is

(a) an allocation for the households $\{(c_{i,j,t})_{j\in[0,1]}, C_t, L_t\}_{t\geq 0}$,

(b) a set of exit decisions for potential entrants and incumbents $\{(\mathbf{1}_{i,t})_{i\in\Lambda_t\cup N_{t-1}}\}_{t\geq 0}$,

(c) an allocation for operating firms $\{(p_{i,t}, y_{i,t}, m_{i,t}, l_{i,p,t}, l_{i,s,t})_{i\in N_t}\}_{t\geq 0}$,

(d) a sequence of aggregate prices $\{W_t, P_{m,t}\}_{t\geq 0}$ and a sequence of sets $\{N_t\}_{t\geq 0}$

such that

1. given (c) and (d), household's allocation in (a) solves their problem in Equation (3.1),
2. given (a) and (d), firms' allocations in (b) and (c) solve their problems in Equation (3.8),
3. labor and matching markets clear:

$$L_t = \int_{i\in N_t} (l_{i,p,t} + l_{i,s,t} + \chi)di, \quad 1 = \int_{i\in N_t} m_{i,t}di$$

4. the set of operating firms, N_t, follows

$$N_t = \{i \in \Lambda_t \cup N_{t-1} : \mathbf{1}_{i,t}\nu_{i,t} = 1\},\ N_{-1} \text{ given.}$$

where $\nu_{i,t}$ is the survival shock for incumbents and is defined to be 1 for entrants at t.

Solution Method We solve the model globally by combining collocation methods and non-stochastic simulation to approximate the distribution of firms. Appendix E provides a description of the recursive formulation of the firm's problem and the computational algorithm that finds the steady state of this economy.

3.5 Efficient Allocation

The endogenous extensive margin of demand creates a new channel for the relationship between size and markups. Not only this new channel affects the joint determination of

size and markup distributions in the equilibrium, but also defines a new Pareto frontier for the economy because the social planner now chooses the distribution of customers across firms. This section characterizes this efficient allocation in our economy.

The Social Planner's Problem Given an initial distribution of productivity, the social planner of this economy maximizes the household's lifetime utility by choosing: (1) which incumbent firms should exit and which potential entrants should enter at each period, (2) how many customers each operating firm should get—which can be achieved either by depreciating their customer base if the firm has too many customers or by launching advertising campaigns if the firm needs to grow—and, finally, (3) how much each operating firm should produce. A formal statement of the planner's problem is included in Appendix D.5.

There are two sources of inefficiencies regarding customer acquisition and the allocation of customers in the equilibrium. First, the planner might choose to allocate customers differently across firms than the equilibrium (misallocation of customers). A second source of inefficiency is the business-stealing externality of advertisement, which leads to an overuse of labor for advertising in the equilibrium. In order to focus on the misallocation of customers, we will abstract away from this second source of inefficiency by restricting the social planner to spend the same amount of aggregate labor for advertisement as in the equilibrium. Therefore, as far as these two sources are concerned, our estimates of inefficiencies in the equilibrium will only reflect the gains from redistributing customers across firms.

The following Lemma shows that restricting the planner to use a certain amount of aggregate labor for advertisement does not restrict their choices for reallocating customers.

Lemma 2. *Any desired distribution of customers across a set of operating firms can be achieved by any strictly positive level of aggregate labor allocated towards advertisement.*

Proof. See Appendix D.6. □

This result follows from the advertisement technology, which requires that returns to advertisement are fully relative in labor allocated towards posting ads.

Given this result, we solve the planner's problem in two steps. First, for any set of operating firms, we characterize the optimal allocation of demand in terms of how many customers each operating firm should get and how much they should produce. Second, we characterize the optimal entry and exit rule that determines the sets of operating firms over time.

Optimal Allocation of Demand Here, we characterize the efficient allocation of customers and demand for a given set of operating firms.

Proposition 3. *Fix a choice for the set of operating firms. Then, under the efficient allocation*

$$q^*_{i,t} = 1, \quad m^*_{i,t} = \frac{z_{i,t}^{\frac{1}{1-\alpha}}}{\int_{i\in N_t} z_{i,t}^{\frac{1}{1-\alpha}} di}.$$

Proof. See Appendix D.7. □

Proposition 3 shows how the planner breaks the link between size and misallocation, and would like all customers to have the *same level of consumption*. Instead, to capitalize on the higher efficiency of more productive firms, the planner gives them *more customers*.[15] This is in contrast to the equilibrium, in which more productive firms have higher sales per customer *and* more customers than other firms (but potentially fewer customers than what is efficient).

Our result in Proposition 3 is also at odds with the trade-off that the social planner faces in conventional models where all firms are assumed to serve the representative consumer. On one hand, the social planner would like more productive firms to produce more to create more *aggregate consumption* (i.e., to equalize marginal product of inputs across firms). However, in those models, since demand comes from the intensive margin, instructing more productive firms to produce more creates dispersion in *relative consumption* across varieties, which is inefficient due to the weak substitutability of goods. Therefore, the social planner has to balance these two opposing forces in choosing the optimal allocation of inputs.

However, in our model, the social planner does not face such a trade-off due to the existence of the extensive margin of demand. The optimal allocation equalizes relative consumption across all customers and, instead, equalizes the marginal product of inputs across firms by giving more customers to more productive firms.

More generally, making the allocation of customers endogenous has two important macroeconomic implications. First, it widens the Pareto frontier of the economy because, in our model, the social planner always has the option to replicate the homogeneous allocation of customers across firms. Second, both the magnitude of losses from misallocation and the distance of the equilibrium allocation from this new frontier, depend on how effective the equilibrium advertising technology is in replicating the efficient allocation of customers rather than the efficient allocation of sales per customer, as is the case in the conventional models.

[15] Note that the allocation of customers does not depend on the initial distribution of matches. This follows from Lemma 2. Since the implementation cost of all distributions is the same for the planner, one can assume without loss of generality that the planner exercises the free disposal of matches in the beginning of every period and re-matches all customers based on firms' new productivities.

Implications for the Number of Firms Here, we derive the planner's policy for entry and exit decisions of firms.

Proposition 4. *Let $v_t^*(z_{i,t})$ denote the social value of a firm with productivity $z_{i,t}$ at time t. Then, this value is given by*

$$v_t^*(z_{i,t}) \equiv \max_{\{(\mathbf{1}_{i,\tau}^*)_{\tau \geq t}\}} \mathbb{E}_t \sum_{\tau=t}^{\infty} (\beta\nu)^{\tau-t} \left(\prod_{h=t}^{\tau} \mathbf{1}_{i,h}^* \right) \left(\frac{C_\tau^*}{C_t^*} \right)^{-\gamma} \left[\underbrace{m_{i,\tau}^* C_\tau^*}_{\text{total sales}} - \underbrace{m_{i,\tau}^* W_\tau^* L_{p,\tau}^*}_{\text{COGS}} - \underbrace{W_\tau^* \chi}_{\text{SGA expenses}} \right],$$

where $(C_\tau^*, L_{p,\tau}^*, W_\tau^*)_{\tau \geq t}$ are the aggregate consumption, aggregate labor allocated to production and the decentralized wage under the planner's solution and $m_{i,t}^*$ is the optimal number of customers given to a firm with productivity $z_{i,t}^*$. The planner keeps a firm in the economy if and only if $v_t^*(z_{i,t}) \geq 0$, captured by the indicator $\mathbf{1}_{i,t}^*$.

Proof. See Appendix D.8. □

To compare this value with the equilibrium value of a firm, we have also separated the terms in the firm's within period value to the equivalent *total sales*, *COGS*, and *SGA expenses*. The terms for the total sales and COGS show that a firm's per period value in production is proportional to the number of customers that the planner allocates towards them. Moreover, the term for the SGA expenses show that in assessing the value of a firm, the planner considers only the amount of overhead labor that she has to spend to operationalize the firm. In particular, the advertisement labor cost of allocating customers does not appear in this value because aggregate advertisement labor is a sunk cost for the planner and, at the margin of keeping a firm, it does not affect her decision.

3.6 Aggregation

Given a set of operating firms, N_t, and an allocation of production inputs $(l_{i,p,t})_{i \in N_t}$ across these firms, we can recast the characterization of the aggregate output and production labor of this economy in the form of an *aggregated production function* as well as an *aggregate markup* that characterizes the wedge between the aggregate marginal product of labor and the wage. In the rest of this section, we derive these aggregate objects and derive decomposition results that allow us to compare the equilibrium and efficient allocations.

Aggregate Production Function We start by deriving the aggregate production function, which can be obtained by defining the *total production labor* as the aggregate amount of labor allocated towards production:

$$L_{p,t} \equiv \int_{i \in N_t} l_{i,p,t} di = \int_{i \in N_t} \left(\frac{C_t m_{i,t} q_{i,t}}{z_{i,t}} \right)^{\alpha^{-1}} di, \tag{3.17}$$

where the second equality follows from the fact that the firm produces to meet its demand as in Equation (3.9).

Defining aggregate output as aggregate consumption, $Y_t \equiv C_t$, and rearranging Equation (3.17), we arrive at the aggregate production function expressed as

$$Y_t = Z_t L_{p,t}^{\alpha}, \tag{3.18}$$

where Z_t, the aggregate TFP, is derived as

$$Z_t \equiv \left[\int_{i \in N_t} \left(\frac{z_{i,t}}{q_{i,t} m_{i,t}} \right)^{-\alpha^{-1}} di \right]^{-\alpha}. \tag{3.19}$$

Aggregate Markup Given an allocation of production inputs and prices among a set of operating firms, $(l_{i,p,t}, p_{i,t})_{i \in N_t}$, we define the aggregate markup, $\mathcal{M}_t$, as the wedge between the aggregate marginal product of labor and the wage W_t. Formally, having derived the aggregate production function in Equation (3.18), the aggregate markup is defined as

$$\mathcal{M}_t \equiv \frac{\partial Y_t / \partial L_{p,t}}{W_t} = \alpha \frac{Y_t}{W_t L_{p,t}}.$$

We can also define the firm level markup as the analog of this wedge for firm i:

$$\mu_{i,t} \equiv \alpha \frac{p_{i,t} y_{i,t}}{W_t l_{i,p,t}},$$

which corresponds to the equilibrium relationship between the markup and the labor share in Equation (3.12)—with the exception that here we are defining this wedge for an arbitrary allocation of inputs and prices. By combining the last two equations, we can then derive the aggregate markup as the production cost-weighted average of firm level markups (as in Edmond et al., 2018):

$$\mathcal{M}_t = \int_{i \in N_t} \omega_{i,t} \mu_{i,t} di, \tag{3.20}$$

where the weight $\omega_{i,t}$ is the *production cost share* of firm i or, as referred to by Baqaee and Farhi (2019), the cost-based Domar weight of firm i:

$$\omega_{i,t} \equiv \frac{W_t l_{i,p,t}}{\int_{i \in N_t} W_t l_{i,p,t}}.$$

Decomposition of Welfare While the planner chooses different distributions of customers and resources across firms, the following proposition shows that all the equilibrium inefficiencies affect the household's welfare only through aggregate objects.

Proposition 5. *For small perturbations around the equilibrium allocation, the welfare losses of the household at a given time t, up to a first order approximation, is given by*

$$\underbrace{\frac{\Delta U_t}{U_{c,t} C_t}}_{\Delta \text{Welfare (C.E.)}} \approx \underbrace{\Delta \ln(Z_t)}_{\Delta TFP} + \underbrace{\alpha(1 - \mathcal{M}_t^{-1}) \Delta \ln(L_{p,t})}_{\Delta \text{Losses from Aggregate Markup}} - \alpha \mathcal{M}_t^{-1} \left[\underbrace{\chi \frac{N_t}{L_{p,t}} \Delta \ln(N_t)}_{\Delta \text{Losses from Entry/Exit}} + \underbrace{\frac{L_{s,t}}{L_{p,t}} \Delta \ln(L_{s,t})}_{\Delta \text{Losses from Advertising}} \right], \tag{3.21}$$

where Z_t is the aggregate TFP in Equation (3.19), $\mathcal{M}_t$ is the equilibrium aggregate (cost-weighted) markup in Equation (3.20), $L_{p,t}$ and $L_{s,t}$ are the aggregate amounts of labor allocated towards production and advertising, and N_t is the equilibrium measure of operating firms.

Proof. See Appendix D.9. □

Equation (3.21) decomposes the consumption-equivalent welfare changes of the household around the equilibrium allocation to four separate terms: (1) allocative and distributional changes that lead to changes in aggregate TFP, (2) losses due to underutilization of labor in the equilibrium that arise from aggregate market power—and demonstrates itself as a wedge between the marginal product of labor and the marginal rate of substitution between consumption and leisure, (3) changes in aggregate labor supply that is allocated towards the overhead costs of operating firms, and (4) changes in aggregate labor supply that is allocated towards advertising in the equilibrium.

Proposition 5 lays out the road map for the rest of our analysis. As we move on to the quantitative part of this study, our main objective is to quantify the importance of the first three channels, while shutting down the fourth by restricting the social planner to use the same amount of aggregate advertising labor as in the equilibrium.[16] While this restriction is a choice on our part and could be easily relaxed, it constitutes a clean benchmark that allows

[16] As we discussed in deriving the efficient allocation, given the business-stealing nature of advertising, all labor that is allocated towards advertising is inefficient from a social perspective.

us to focus on quantifying the main channel of interest in our analysis, the misallocation of demand across firms.

3.7 Discussion of Assumptions

Before moving on to our quantitative analysis, here we discuss our main model assumptions.

Mechanisms for Customer Acquisition In Equation (3.5) we have assumed a law of motion for customers that rules out (1) customer acquisition through prices, and (2) customer retention through firm decisions (i.e., an endogenous separation rate).

The assumption that firms cannot attract *new* customers by lowering their prices is motivated by our empirical results in Table 3 that markups are not correlated with the size of firms' customer bases conditional on sales per customer. Fitzgerald et al. (2016) also conclude that firms do not manipulate prices to shift demand by documenting that after firms enter into a new market their markups remain the same while their quantities grow. Instead, based on our findings in Table 4, we model customer acquisition through advertising-like activities (as in, e.g., Arkolakis, 2010; Drozd and Nosal, 2012; Sedláček and Sterk, 2017). Table 4 also shows that SGA expenses do not covary with the retention of old customers, which is why we assume that the separation rate of customers is exogenous to firm decisions.

Therefore, in our model, customer acquisition can have the interpretation of a process through which potential customers become "aware" of a product or a firm (as in Perla, 2019). Once they are aware of the product, however, the price affects their demand only in the intensive margin and it does not affect their decision to leave the firm.

The Relative Nature of Advertising We have made the assumption that advertising, while affecting the distribution of customers across firms, does not increase the total number of customers that buy from an industry—i.e., it has a business-stealing nature (as in Drozd and Nosal, 2012; Einav et al., 2020). This assumption is supported by evidence from the marketing and IO literature.[17] For instance, studying the publishing sector, Garthwaite (2014) writes, "[book] endorsements are found to be a business-stealing form of advertising that raises title level sales without expanding the market size." Similarly, for prescription drugs, Sinkinson and Starc (2019) find a positive own-elasticity of revenue to advertising, but a negative and similar in size negative elasticity of revenue with respect to rival advertising. For consumer packaged goods, Hartmann and Klapper (2018) find that advertising increases revenue per household, but when two major brands advertise together, this revenue

[17] See Bagwell (2007) for a discussion.

is lost. More specifically, in our model, this assumption ensures that independent of how much firms spend on advertising within an industry, they cannot create new customers.[18]

4 Quantitative Analysis

To quantify the implications of customer acquisition for the efficiency losses from market power, we calibrate the steady state of the model using the Simulated Method of Moments and matching several micro- and macro-moments related to firm dynamics in the US economy in 2012.

4.1 Calibration Strategy

To provide an overview of our calibration strategy, the new and the most relevant parameter to calibrate is ϕ, the returns to scale in advertising—which disciplines the strength of the relationship between a firm's size and market power, by determining the size of its customer base.[19] Moreover, as we show in Appendix B.2, the composition of firms' costs exhibits a significant size profile based on firms' sales. Therefore, it is also important to have a good fit for the sales distribution in the economy. Ideally, we would combine aggregate data on the size distribution of firms with aggregate data on firms' cost structures. The fact that the latter are only available from Compustat, which includes only a subset of firms in the economy, poses a challenge. We address this challenge in the following way. Whenever possible, we calibrate the model to the aggregate US economy in 2012 by matching moments from the Business Dynamics Statistics (BDS) and Statistics of US Businesses (SUSB) provided by the Census Bureau. When matching moments regarding firms' costs, we apply a filter in the simulated data to account for the selection into Compustat based on size and age. In the remainder of this section, we provide a detailed description of our calibration strategy.

Fixed Parameters We set the length of a period to one year. Panel A of Table 5 presents the set of parameters that are externally fixed. We set the subjective discount factor β to match an annual interest rate of 4%. The elasticity of intertemporal substitution γ is set to 2. We set the inverse of the Frisch elasticity of labor supply to $\psi = 1$ and the labor coefficient in the production function to $\alpha = 0.64$. In the calibration exercise, we normalize the measure

[18] It is also important to note that in our characterization of the efficient allocation, even the planner is subject to this restriction: she cannot create more customers. This ensures that our measurement of welfare gains does not hinge on more relaxed feasibility constraints for the planner and only comes from reallocation of customers.

[19] Note that our model does not require firms to spend on customer acquisition and grow through the extensive margin of demand. This is because $\lim_{\phi \to 0} l^{\phi}_{i,s,t} = 1$ and $\lim_{\phi \to 0} Wl_{i,s,t} = 0$. Thus, our model nests the conventional model with exogenous customer bases as a special case.

of potential entrants λ and the disutility of labor supply ξ to generate a steady-state output of $Y = 1$ and wage of $W = 1$.

We set the retention rate of customers to $1 - \delta = 0.72$, which corresponds to the repurchase probability in the Nielsen-GS1 matched dataset in 2012.[20] Although we use Nielsen-GS1 matched data, which is limited to the consumer packaged goods sector, the repurchasing probability is similar in other industries based on evidence from the marketing literature. For example, the repurchase probability is 0.7 in the automotive industry based on survey data used in Mittal and Kamakura (2001). According to Bolton, Kannan and Bramlett (2000), the loyalty program member share is 0.693 and the cancellation probability is 0.187 for the financial service industry. Finally, Bornstein (2018) estimates an annual retention probability of 0.85 for the top two largest firms in each product category from the Nielsen data. If we also restrict the sample to the top two firms, our retention measure increases to 0.84.

Calibrated Parameters We jointly calibrate the remaining 8 parameters by the simulated method of moments (SMM).[21] These parameters can be grouped in three sets, those shaping firms' cost structure (ϕ and χ), their demand (σ, and η), and their life cycle and shock structure ($\rho_z, \sigma_z, \bar{z}_{ent}$ and ν). Although these parameters are jointly identified by all moments, we provide below a discussion of which moment should intuitively be more relevant to identify each parameter. We formalize this discussion in Appendix F by analyzing the local elasticities of model moments with respect to each parameter and the sensitivity measure developed by Andrews, Gentzkow and Shapiro (2017).

To calibrate the overhead cost χ, we target the cross-sectional average COGS-to-OPEX ratio from Compustat. The model counterpart of this ratio for firm i is

$$\frac{W_t l_{i,p,t}}{W_t l_{i,p,t} + W_t \left(l_{i,s,t} + \chi\right)} \equiv \frac{COGS_{i,t}}{COGS_{i,t} + SGA_{i,t}}.$$

Intuitively, a larger fixed cost χ, ceteris paribus, should increase a firm's total costs and drive

[20] More specifically, define $\text{Sales}_{i,g,t}$ as the total expenditure of (projection-factor adjusted) households who purchase products made by firm i in group g at time t. Define the probability of repurchasing firm's products as $s_{i,g,t} = \frac{\text{Sales}_{i,g,t-1,t}}{\text{Sale}_{i,g,t-1}}$, where $\text{Sale}_{i,g,t-1,t}$ is the total expenditure of (projection-factor adjusted) households who purchase products made by firm i in group g in both periods $t-1$ and t. Then, we take a weighted average of $s_{i,g,t}$ across firms and groups, where the weights are the expenditure in firm-group bins across all years.

[21] More specifically, we calibrate the model by choosing a set of parameters $\mathcal{P}$ that minimizes the SMM objective function

$$\left(\frac{\boldsymbol{m}_m(\mathcal{P})}{\boldsymbol{m}_d} - 1\right)' \boldsymbol{W} \left(\frac{\boldsymbol{m}_m(\mathcal{P})}{\boldsymbol{m}_d} - 1\right),$$

where $\boldsymbol{m}_m$ and $\boldsymbol{m}_d$ are a vector of model simulated moments and data moments, respectively, and $\boldsymbol{W}$ is a diagonal matrix. Appendix E.2 provides the computational details of the calibration exercise.

down this ratio.

To identify the elasticity ϕ, we exploit the observed relationship between SGA and sales in Compustat. The following proposition illustrates the source of identification in the special case of the model with $\delta = 1$, which admits a closed-form solution of the firm's optimal spending in customer acquisition.

Proposition 6. *Suppose $\delta = 1$. Then, the total $SGA_{i,t}$ expenses of a firm can be decomposed into a fixed $(SGAF_{i,t})$ and a variable $(SGAV_{i,t})$ component:*

$$\begin{aligned} SGA_{i,t} &= SGAF_{i,t} + SGAV_{i,t} \\ &= W_t\chi + \phi Sales_{i,t} - \frac{\phi}{\alpha} COGS_{i,t} \end{aligned} \tag{4.1}$$

Proof. See Appendix D.10. □

Equation (4.1) is obtained from the firm's optimality condition regarding customer acquisition. The firm acquires customers up to the point where the marginal cost of an additional customer equals the marginal benefit, which equals profits from those marginal sales. In this special case, ϕ is identified by the relationship between $SGA_{i,t}$ expenses and sales, conditional on $COGS_{i,t}$ and time fixed-effects. For the relevant case with $\delta < 1$, Appendix F shows a high sensitivity of ϕ to the same relationship. Thus, we calibrate ϕ to match the coefficient on $Sales_{i,t}$ from an OLS regression of Equation (4.1) using data from Compustat. In the model, we compute the moments related to firms' cost structures after accounting for selection into Compustat with two filters based on firms' age and size. That is, to compute these moments we restrict the simulated sample of firms to those that are at least 7 years old, as in Ottonello and Winberry (2018), and have sales above 19% of the average sales in the simulated economy (which corresponds to the ratio of the 5th percentile of the sales distribution in Compustat to average sales in SUSB).[22]

To calibrate the parameters shaping firms' demand, we set the elasticity of substitution σ to match a COGS-weighted average markup of 1.25 computed from Compustat. This follows from the aggregation result that shows that the relevant aggregate markup corresponds to the production cost-weighted average markup. Second, following Edmond et al. (2018), the super-elasticity of demand η is pinned down by the relationship between a firm's relative average revenue productivity of labor and its relative sales. In a model without customer acquisition, the revenue productivity of labor $p_{i,t}y_{i,t}/W_t l_{i,p,t}$ is directly proportional to the production markup $\mu_{i,t}$. The following proposition shows that a similar relationship

[22] Average firm sales in the 2012 US economy were USD5.7 million (SUSB) and the 5th percentile of the sales distribution in Compustat was USD1.06 million.

holds in a special case of the model.[23] Appendix F shows that the super-elasticity η is sensitive to this relationship in the general model as well.

Proposition 7. *Suppose $\delta = 1$. Then, a firm's average revenue productivity of labor is given by*

$$\frac{p_{i,t}y_{i,t}}{W_t(l_{i,p,t} + l_{i,s,t})} = \frac{\mu_{i,t}}{\alpha + \phi(\mu_{i,t} - 1)}$$

which is strictly increasing in the production markup $\mu_{i,t}$ if and only if $\alpha > \phi$.

Proof. See Appendix D.11. □

When $\eta = 0$, markups and the revenue productivity of labor are constant and independent of sales. When $\eta > 0$, both markups and sales are increasing in productivity. Therefore, the relationship between labor productivity and sales is informative about η, holding the other parameters fixed. We summarize this relationship with the regression coefficient of a sales-weighted OLS regression of relative revenue productivity of labor on relative sales of 0.036 for firms with relative sales greater than 1, as reported by Edmond et al. (2018) and computed using aggregate data from SUSB.[24]

Finally, the parameters of the AR(1) productivity process for incumbent firms, σ_z and ρ_z, are set to match a standard deviation of annual employment growth of 0.415 from Elsby and Michaels (2013) and the unweighted distribution of within-industry relative sales from Edmond et al. (2018). The mean of the productivity distribution of entrants $\bar{z}_{ent}$ is set to match the fact that old firms (those older than 11 years) are on average six times larger in terms of employment than 1-year old firms (BDS). The exogenous separation probability ν is calibrated to match an average exit rate of 7.3% (BDS).

Results The set of calibrated parameters is shown in Panel B of Table 5. The process for the productivity shock is quite persistent and volatile, although in line with estimates from Lee and Mukoyama (2015). The calibrated elasticity and super-elasticity of demand are 6.49 and 4.95, respectively, which are close to values used and estimated in the literature

[23] The inverse of this relationship—a firm's labor share—is the inverse-markup-weighted average of the returns to scale in different uses of labor. A similar relationship appears in Kaplan and Zoch (2020).

[24] In our definition of model revenue productivity of labor, we include the variable component of SGA (l_s) but not the fixed component of SGA (χ). The former is due to the fact that the SUSB reports information on the total wage bill across firms in a size group, without distinguishing between types of labor (e.g., production and advertising labor). The decision not to include χ is due to the fact that part of overhead costs are, in reality, not associated with labor costs (e.g., rent) and thus not included in the wage bill reported by SUSB. Ideally, we would use data on the subcomponents of SGA expenses in Compustat to compute the share of labor costs within SGA expenses. Unfortunately, a full disaggregation of SGA expenses is not available. To alleviate concerns about this choice, note that we target a moment based on a sample of relatively large firms (those with relative sales greater than 1), for which arguably the fixed overhead cost represents a smaller fraction of total costs.

Table 5: Model Parameters

Parameter	Description	Value
	Panel A: Fixed Parameters	
β	Annual discount factor	0.960
γ	Elast. of intertemporal substitution	2.000
ψ	Frisch elasticity	1.000
α	Decreasing returns to scale	0.640
δ	Prob. of losing customer	0.280
	Panel B: Calibrated Parameters	
ϕ	Elasticity matching function	0.533
χ	Overhead cost	0.307
σ	Avg. elasticity of substitution	6.490
η	Superelasticity	4.956
ν	Exog. survival probability	0.964
ρ_z	Persistence of productivity shock	0.973
σ_z	SD of productivity shock	0.218
$\bar{z}_{ent}$	Mean productivity of entrants	-1.453
λ	Mass of entrants	0.137
ξ	Disutility of labor supply	1.981

Notes: This table shows the calibration of the model. Panel A contains parameters externally chosen. Panel B contains parameters internally calibrated to match moments presented in Table 6 and Figure 3.

(see e.g., Gopinath and Itskhoki (2010); Nakamura and Zerom (2010)). Finally, note that the calibrated value for the elasticity of the matching function $\phi = 0.533$ is close to a model-generated regression coefficient of 0.474. This similarity lends support to the identification argument provided in Proposition 6.

Table 6 and Figure 3 show the targeted moments and their model counterparts. Overall, the model closely matches the targets. The model is able to reproduce the average cost structure very well, but it slightly under-predicts the relationship between SGA and sales. The model matches well the cost-weighted average production markup, and is able to generate a similar relationship between revenue productivity of labor and sales. Figure 3 shows that the model approximates well the sales distribution of firms. For example, in the data 33% of firms have sales that are lower than 10% of the average sales in the economy and 1% of firms have sales that are larger than 10 times the average sales. In the model, these shares are 25% and 1.5%. The Figure also shows that the model is able to replicate the relative size of old firms, 6.07 in the data and 6.4 in the model. In Appendix F, we show the data

Table 6: Targeted Moments

Moment	Data	Model
Slope SGA on sales	0.492	0.474
Avg. COGS-to-OPEX ratio	0.660	0.669
Avg. cost-weighted production markup	1.250	1.275
Slope labor prod. on sales	0.036	0.033
Avg. exit rate	0.073	0.071
SD. employment growth	0.416	0.447

Notes: This table shows the set of moments targeted in the calibration of the model. Slope SGA on Sales refers to the OLS coefficient of the regression $SGA_{i,t} = c + \beta Sales_{i,t} + \psi COGS_{i,t} + \varepsilon_{i,t}$. Avg. COGS-to-OPEX ratio refers to the average of the ratio across firms. Avg. cost-weighted production markup corresponds to the COGS-weighted average markup from Edmond et al. (2018). These moments were computed using data from Compustat in 2012. Slope labor prod. on sales corresponds to the OLS coefficient of the sales-weighted regression of relative revenue labor productivity on relative sales from Edmond et al. (2018), restricting the sample of firms with relative sales above one. This moment was computed using data from the SUSB in 2012. The average exit rate was obtained from the BDS in 2012. The standard deviation of annual employment growth for continuing establishments is obtained from Elsby and Michaels (2013). Growth rate of variable x is computed as in Davis and Haltiwanger (1992): $(x_{i,t} - x_{i,t-1})/(0.5(x_{i,t} + x_{i,t-1}))$. The last column shows the model counterparts of each moment, which were obtained by simulating a panel of firms and computing each moment with the simulated data. In the model, we account for selection into Compustat by restricting the simulated sample of firms to those that are at least 7 years old and have sales above 19% of the average sales in the simulated economy (which corresponds to the ratio of the 5th percentile of the sales distribution in Compustat to the average sales in SUSB).

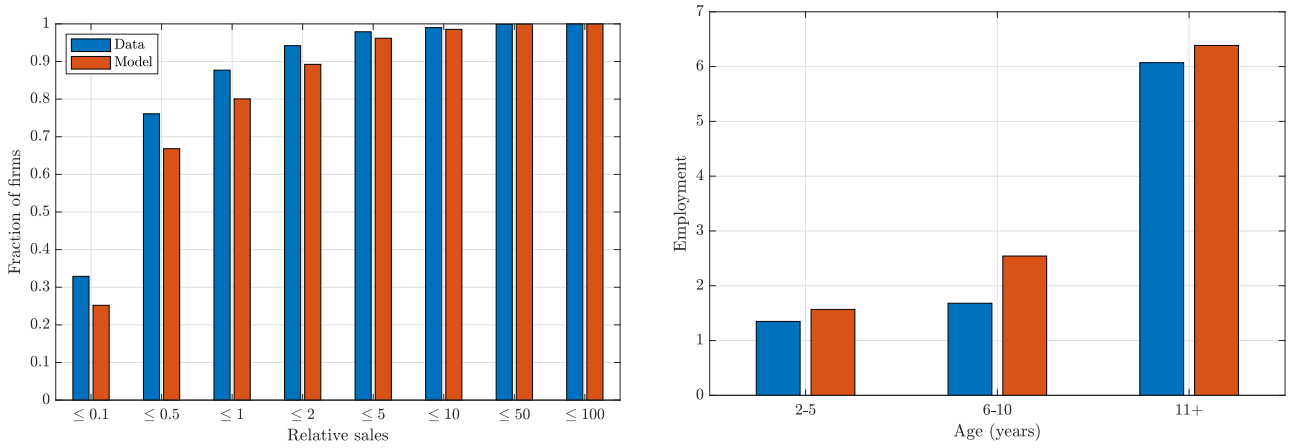
and model relationships between average labor productivity and SGA with sales, and the average COGS-to-OPEX ratio by firm age and size. Although in the calibration exercise we targeted specific moments of these relationships, the model matches the data patterns more broadly.

4.2 Model Validation

Before proceeding with the main quantitative analysis, we provide over-identifying tests of the calibrated model regarding its ability to match relevant untargeted moments. First, we show that the model is able to generate firm dynamics similar to those observed in the data. Second, we test the model's predictions regarding the co-movement between a firm's production markup, average sales per customer and the size of the customer base.

Firm Dynamics Figure 4 shows two model moments that were not explicitly targeted during the calibration exercise: the average exit rate by age and the average employment growth by age. As Panel A shows, the average exit rate is decreasing in the firm's age, as in the data (see e.g., Haltiwanger, Jarmin and Miranda, 2013). The fact that entrants enter the

Figure 3: Model Fit: Firm Size



(a) Distribution of Sales

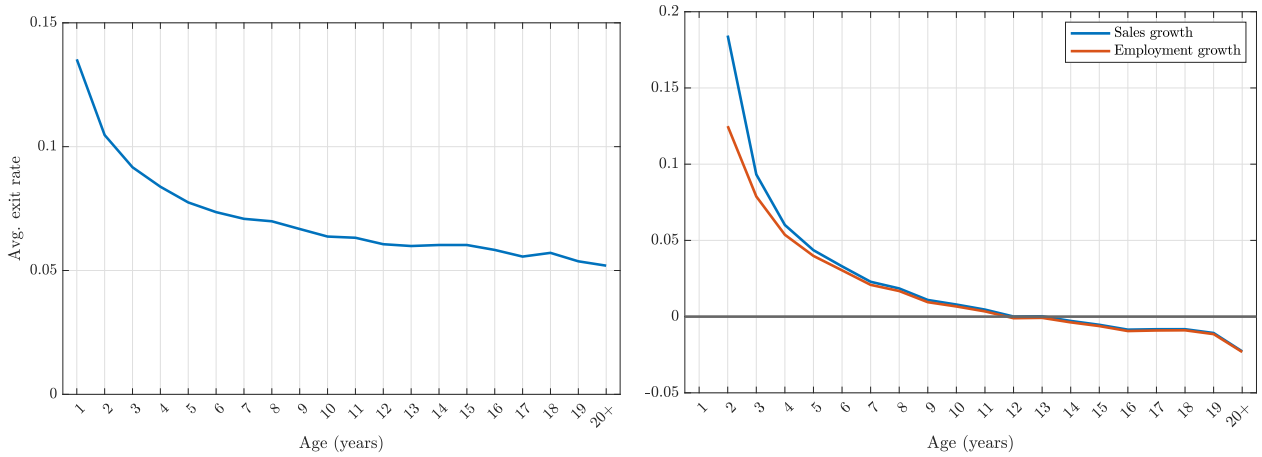
(b) Avg. Size by Age

Notes: This figure shows moments targeted in the calibration of the model. Panel (a) shows the model fit of the distribution of relative sales. The distribution of relative sales is obtained from the SUSB in 2012. Panel (b) shows the model fit of average employment, relative to 1 year-old firms, by firm age. Average firm employment by age group was obtained from BDS in 2012. In the calibration exercise, we only target the relative size of firms older than 10 years.

economy with lower average productivity and no customer base makes young firms more likely to exit when faced with negative productivity shocks due to the presence of overhead costs. As firms grow larger, their larger customer base and higher productivity allows them to absorb negative productivity shocks without forcing them to exit. It is important to note that in the firm dynamics literature, the decreasing profile of the average exit rate by age is typically used as a target to indirectly calibrate the size of the overhead cost χ. Here, we took another route by directly matching the observed structure of firms' costs. The fact that the model is able to match well the profile of exit rates provides additional support to the modeling and split of the firm's cost structure. Panel B plots decreasing profiles of average employment and sales growth as a function of firms' age. Both patterns are consistent with the empirical evidence in Haltiwanger et al. (2013). For example, in the data, the average net employment growth rate of 1-2 year-old and 7-8 year-old continuing firms is close to 12% and 2.5%, respectively. In the model, the average employment growth rates (computed as in Davis, Haltiwanger, Schuh et al. (1998)) for 2- and 7-year-old firms are 12.5% and 2.1%, respectively.

Drivers of Firm Growth We have previously documented that, in the data, the major source of cross-firm differences in sales is the size of their customer bases. Here, we verify the extent to which the model is able to quantitatively match this fact. Table 7 compares the variance decomposition of log sales in the model with the decomposition from the data. In the data, differences in log average sales per customer account for 11.4% of the variance of

Figure 4: Exit and Growth by Age



(a) Exit Rate by Age

(b) Avg. Growth by Age

log sales. The model closely matches this fact, with a fraction of 15.2%. Also, in the model the largest contributor to the dispersion in log sales is the variance of the log number of customers, as in the data. However, since differences across firms are ultimately driven by only one source of heterogeneity (i.e., the productivity shocks) the model naturally over-predicts the size of the covariance term.

Table 7: Sources of Sales Dispersion across Firms

	Var(ln sales per customer)	Var(ln n. of customers)	Covariance
Data	11.44	80.66	7.90
Model	15.17	47.54	37.29

Notes: This tables provides a variance decomposition of firms' log sales. The first column reports the variance of the log sales per customer, $var\left(\ln p_{i,t}\mathrm{Y}'(p_{i,t}/D_t)\right)$, relative to the overall variance of log sales. The second column reports the relative variance of the log number of customers, $var\left(\ln m_{i,t}\right)$. The last column reports the covariance between both terms, $cov\left(\ln m_{i,t}, \ln p_{i,t}\mathrm{Y}'(p_{i,t}/D_t)\right)$. The first row reports the results obtained from the Nielsen Homescan Panel. Sales and the number of customers are adjusted with household sample weights. The second row reports the results obtained from model simulated data.

Relatedly, we have shown that, despite not being the main driver of sales growth, average sales per customer are strongly associated with market power (see Table 3). In this section, we show that our model is able to reproduce this fact quantitatively despite not being a direct target in the calibration. For this, we regress simulated firms' markups on sales per customer, the size of their customer bases and time fixed effects:

$$\ln(\mu_{i,t}) = \theta_0 \ln p_{i,t}q_{i,t} + \theta_1 \ln m_{i,t} + \kappa_t + \varepsilon_{i,t}.$$

Table 8 presents the results. The data show a statistically significant relationship between markups and sales per customer, and an economically and statistically insignificant relationship between markups and the size of the customer base. The model matches these facts fairly well. The model predicts that 1% higher average sales per customer are associated with 0.11% higher markups. This point estimate is in between the baseline estimate of 0.06 reported in Table 3 and the estimate of 0.187 reported in the additional analysis in Appendix B.1.3. On the other hand, a 1% increase in the size of the customer base increases markups by only 0.02%.[25] Therefore, the model captures the differential roles of intensive and extensive margins of demand on firms' markups, as we document in the data.

Table 8: Sources of Dispersion in Sales and Markups

	$\ln \text{Markup}_{it}$	
	Data	Model
$\ln p_{it} q_{it}$	0.060^{***}	0.111
	(0.024)	
$\ln m_{it}$	0.003	0.022
	(0.007)	
Observations	2433	
R^2	0.338	0.869
Year FE	✓	✓
SIC FE	✓	

Notes: This table reports the results of an OLS regression of a firm's log markup ($\ln(\mu_{i,t})$) on log sales per customer ($\ln p_{i,t} q_{i,t}$) and log size of the customer base ($\ln m_{i,t}$). Column (1) reproduces the empirical estimates from Table 3. Column (2) reports estimates based on model-simulated data. The model-simulated panel is restricted to mimic selection into Compustat (see Section 4 for details). In the model, we do not include SIC FE as we model a single "representative" industry.

5 The Role of Endogenous Customer Acquisition

This section investigates the role of endogenous customer acquisition in (1) directly shaping a firm's optimal choices and (2) indirectly determining the aggregate properties of the equilibrium. To do so, we compare the equilibrium allocation of the calibrated model (labeled as "Baseline" from hereon) with the allocation under an alternative model where firms receive a fixed and equal number of total customers in each period, without having to spend

[25] In our model the size of the customer base is a demand shifter and does not directly affect the elasticity of demand. The only reason that this regression coefficient is not 0 in the simulated data is the nonlinear nature of the relationship between these variables.

any resources (labeled as "Restricted" from hereon). We present this comparison in two settings. First, to illustrate the mechanisms at play, we provide *comparative statics* by assuming the same set of parameters across both models, while shutting down endogenous customer acquisition in the Restricted one. Second, to study the quantitative implications, we re-calibrate the Restricted model to the same set of empirical moments minus the one that pins down returns to advertising ϕ, and compare the implied distribution of markups across the two models.

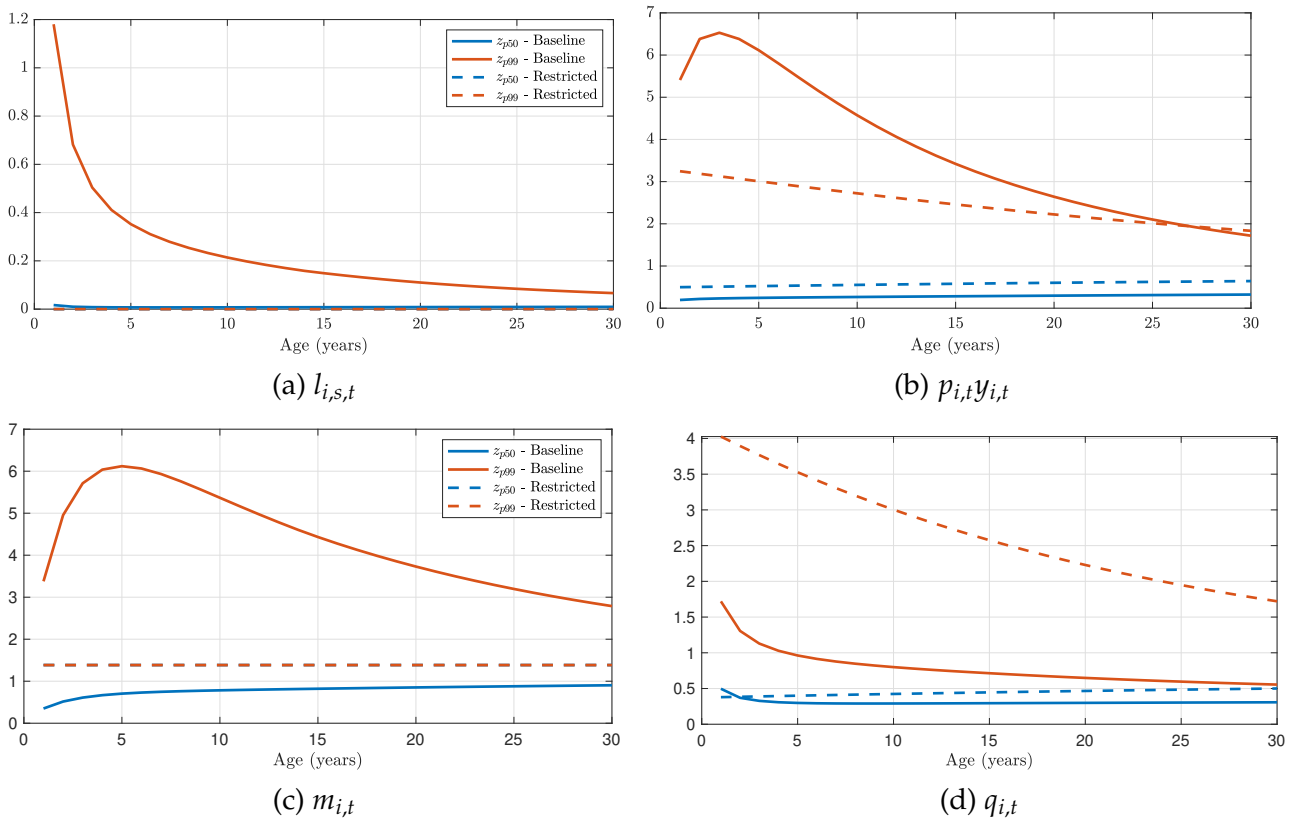
5.1 Comparative Statics

Implications for Concentration Figure 5 plots the firm dynamics of two entrants that start with productivities equal to the 50th and 99th percentile of the productivity distribution of entrants. After these initial draws, productivity follows the AR(1) process without any further shocks. Panel (a) shows that in the Baseline model, firms front load their efforts to acquire customers when young with more productive firms spending more on advertising labor, $l_{i,s,t}$. Panel (b) shows that the sales of the more productive firm are larger than the ones of the less productive firm, and also that sales increase initially due to a growing stock of customers but eventually decline due to the mean-reversion of its productivity.

How do firms grow their sales? Panels (c) and (d) of Figure 5 plot the dynamics of the number of customers and sales per customer, respectively. Both firms build their customer base gradually over time due to decreasing returns to advertising. However, sales per customer are higher when firms are younger due to mean-reverting productivity and decreasing marginal product of labor in production. Thus, over time, firms shift their sales strategy from selling more to few customers, to selling less to more customers.

How does *endogenous* customer acquisition affect these dynamics? Relative to the Restricted model, the more productive firm is able to achieve higher sales by selling less per customer, but accumulating more than twice as many customers in the first year. Since more productive firms charge lower prices but higher markups, profits per marginal customer are increasing in productivity, which induces the more productive firm to accumulate customers more rapidly. Given a fixed stock of customers, this can only be possible if firms with lower productivity accumulate fewer customers, relative to the Restricted model. Thus, endogenous customer acquisition increases the dispersion across firms of the number of customers and decreases the dispersion of sales per customer. Table 9 shows that the former effect dominates and overall concentration of sales increases: while in the Restricted model, the 5% of the largest firms capture 17% of sales, in our Baseline model they capture 50% of sales.

Figure 5: Average Firm Dynamics



(a) $l_{i,s,t}$ (b) $p_{i,t}y_{i,t}$

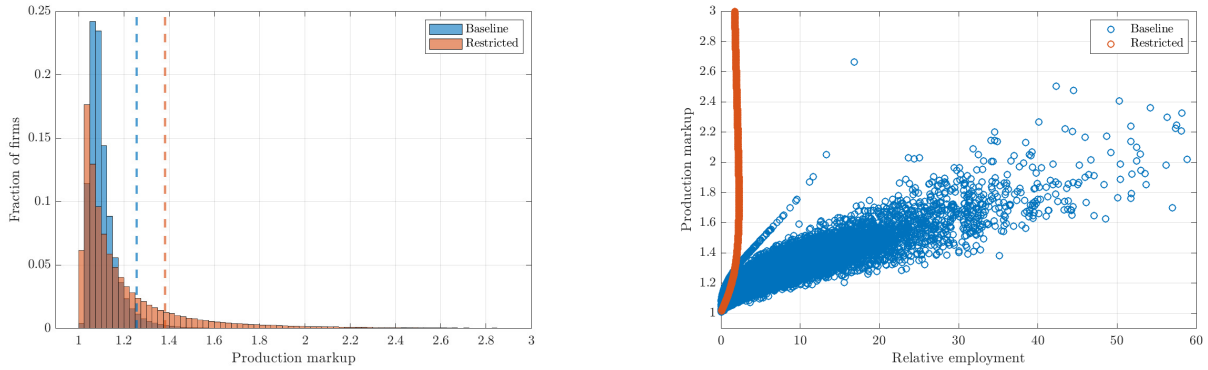
(c) $m_{i,t}$ (d) $q_{i,t}$

Notes: This figure plots the firm dynamics of two new firms that start with zero customer base and productivities equal to the 50th and 99th percentile of the distribution of productivities among entrants. After this initial draw, firms follow the AR(1) productivity process without any further shock. Solid lines correspond to the "baseline" model. Dashed lines correspond to the "restricted" model with an exogenous customer base ($m_{i,t} = 1/N_t$). For the latter, we compute the general equilibrium using the calibrated parameters of the baseline model. Panels (a)-(d) plot the evolution of labor devoted to customer acquisition ($l_{i,s,t}$), sales ($p_{i,t}y_{i,t}$), customer base ($m_{i,t}$) and output per customer ($q_{i,t}$), respectively.

Implications for Market Power Table 9 shows that despite higher concentration, the Baseline model features a lower aggregate markup, 1.26 as opposed to 1.38 in the Restricted model. To understand the sources of this difference, Figure 6 shows the histogram of markups and the scatter plot between markups and relative employment (the weights used in the construction of the aggregate markup) across the two models. These figures illustrate two forces. On one hand, in the model with endogenous customer acquisition the distribution of markups is more concentrated. That is, high productivity firms charge lower markups relative to firms with similar productivity in the Restricted model. On the other hand, in our model those high productivity firms account for a larger fraction of total employment.

The following decomposition of the difference in aggregate markups, which follows

Figure 6: Customer Acquisition and Market Power



(a) Distribution of Markups

(b) Markups vs. Relative Employment

Notes: Panel (a) plots the distribution of production markups in the Baseline and Restricted models. The vertical dashed lines show the average cost-weighted production markup in each model. Panel (b) shows the scatter plot of relative employment ($l_{i,s,t}/(L_{s,t}/N_t)$) and production markups $\mu_{i,t}$. The restricted model refers to the model with an exogenous customer base ($m_{i,t} = 1/N_t$).

from Equation (3.20), quantifies each force:

$$\underbrace{\ln(\mathcal{M}_t) - \ln(\mathcal{M}_t^R)}_{-9.71\%} \approx \underbrace{\int_{i\in N_t} (\omega_{i,t} - \omega_{i,t}^R)\ln(\mu_{i,t})di}_{\Delta \text{ Distribution: } 9.00\%} + \underbrace{\int_{i\in N_t^R} \omega_{i,t}^R \left(\ln(\mu_{i,t}) - \ln(\mu_{i,t}^R)\right) di}_{\Delta \text{ Market power: -18.71\%}},$$

where the superscript R denotes distributions and allocations in the restricted model.[26] The first term (denoted "Δ Distribution") captures the contribution of differences in the distribution of relative employment across firms while keeping firms' markups fixed at their level in the Baseline model. The second term (denoted "Δ Market power") captures the contribution of differences in markups across models while keeping the distribution of relative employment fixed at its distribution in the Restricted model. As more productive firms charge higher markups, switching the distribution of relative employment from the Restricted to the Baseline model would *increase* the average markup by 9pp. However, the contribution of lower markups in the Baseline model *reduces* the aggregate markup by 18.7pp, so the aggregate markup ends up being smaller by 9.71pp. To summarize, why does the baseline model feature much higher concentration but a lower aggregate markup? Because in the baseline model firms grow through larger customer bases $m_{i,t}$, rather than higher average sales per customer $p_{i,t}q_{i,t}$, which reduces their market power (but increases their lifetime

[26] While the equation presents the decomposition based on its approximation for expositional purposes, the numbers we present are computed based on the exact decomposition.

profits).

Aggregate Implications Beyond market power, endogenous customer acquisition also affects other aggregate outcomes by allowing consumers to be concentrated among high productivity firms. First, relative to the Restricted model, it allocates more of the economy's resources towards more productive firms, which reflects itself in higher aggregate TFP derived in Equation (3.19). Table 9 shows that shutting down endogenous customer acquisition in our model reduces aggregate TFP by 28 percent. This is because in the Baseline model more productive firms utilize a higher share of production resources to meet the demand from their larger customer bases.

Second, the concentration of customers among more productive firms leaves fewer customers for firms at the bottom of the productivity distribution and brings them closer to the exit threshold. Therefore, when we restrict customer bases to be equal across firms, the equilibrium number of firms increases by 65 percent. This additional inflow of firms comes from less productive firms that can now generate positive discounted profits due to a larger (exogenous) customer base.

The higher number of firms in the Restricted also leads to higher demand for employment. Table 9 shows that in the Restricted model total employment is 8 percent larger that in the Baseline model (despite the fact that there is no spending in customer acquisition in the former). Part of this difference stems from the larger number of firms that requires larger spending in overhead costs. The table also shows that aggregate *production* labor is also 6.3 percent larger. This is the result of income effects that increases labor supply due to lower aggregate consumption. Finally, the changes in aggregate TFP and production labor, together, lead to an overall 24 percent decline in aggregate output.

5.2 Recalibrating the Restricted Model

The motivating question for this section is: how would the implied distribution of markups differ if we calibrated both models to the same distribution of firm size? To answer this question, we recalibrate the Restricted model to the same set of moments that we targeted for the Baseline model, dropping the moment on the relationship between SGA expenses and sales that was used to pin down returns to advertising, ϕ. The results of this calibration as well as results for goodness of fit are reported in Appendix G.

The main observation is that by matching the same distribution of sales as in the Baseline model (Figure G.1), the Restricted model assigns a much lower calibrated value to the parameter determining the super-elasticity of demand (η/σ), similar to the one reported in Edmond et al. (2018). The reason for this lower value is that the model is forced to gener-

Table 9: Aggregate Effects of Customer Acquisition

	Baseline Model	**Restricted Model**
TFP		-27.9
Output		-23.9
Number of firms		65.1
Employment		7.9
Production		6.3
Agg. markup	1.26	1.38
Top 5% sales share	0.50	0.17

Notes: The table reports equilibrium aggregates in the baseline and restricted versions of the model. The restricted model refers to the model with an exogenous customer base ($m_{i,t} = 1/N_t$). The second column reports percentage differences with respect to aggregates in the baseline model, with the exception of the aggregate markup and the top 5% sales share, which are reported in levels.

ate the wide distribution of sales using only the *intensive margin* of demand, i.e., it needs to distribute all firms on the same demand curve. However, since higher η diminishes the possibility of selling high quantities to the representative consumer (by generating higher elasticities at higher quantities), the model needs a low value of η to match the dispersion of sales in the data. It is important to note that the Baseline model is not constrained by a high value of η as it can match the dispersion of sales by placing firms on parallel demand curves using the *extensive margin* of demand.

How does shutting down endogenous customer acquisition distort our interpretation of the data through the lens of the model? As shown in Figure G.2, because of the much lower value of η, the recalibrated Restricted model generates much lower markup dispersion with a higher unconditional average markup across firms.[27] Two observations follow. First, the Restricted model overestimates markups for small firms and underestimates markups for large firms. Second, by generating a higher markup dispersion than the Restricted model, the Baseline model implies much higher welfare losses due to misallocation, which previews the main difference between our results and those in Edmond et al. (2018), and brings us to our last exercise.

[27] Recall that with $\eta = 0$, the Kimball aggregator converges to the CES aggregator and markup dispersion is zero.

6 Quantifying the Efficient Allocation

The objective of this section is twofold. First, we present the differences between the equilibrium and efficient allocations and, in particular, we quantify the gains under the efficient allocation of demand. In doing so, we also revisit our ex-ante decomposition of welfare in Proposition 5 and quantify the contribution of each of the three channels (TFP, Aggregate Markups and Overhead costs).

Second, in order to isolate the role of endogenous customer acquisition, we repeat our first exercise for counterfactual values of ϕ—the parameter that governs the proximity of the equilibrium allocation of customers to the Pareto frontier—and study how welfare losses change once the equilibrium distribution of customers gets closer to the efficient allocation.

Gains in Welfare We start by quantifying the three channels of welfare gains from Proposition 5 in our calibrated model:

$$\underbrace{\frac{\Delta U_t}{U_{c,t}C_t}}_{\Delta\text{Welfare (C.E.) = 13.6\%}} \approx \underbrace{\Delta \ln(Z_t)}_{\text{TFP gains = 10.8\%}} \underbrace{-\alpha \mathcal{M}_t^{-1} \chi \frac{N_t}{L_{p,t}} \Delta \ln(N_t)}_{\text{Gains from Entry/Exit = 1.6\%}} \underbrace{+\alpha(1-\mathcal{M}_t^{-1})\Delta \ln(L_{p,t})}_{\text{Losses from Underutilization of Labor = 0.78\%}}$$

There are two main takeaways from this decomposition: (1) the consumption-equivalent welfare gains of the household under the efficient allocation are substantial and quantified at 13.6%, (2) the majority of this gain is coming from the efficiency gains in aggregate TFP under the planner's allocation, quantified at 10.8% higher than the equilibrium TFP. In addition to this substantial gain in TFP, the planner is also able to generate 1.6% higher welfare by reducing the amount of labor allocated towards the overhead cost of operating firms, and 0.78% higher welfare by correcting for the underutilization of labor due to aggregate market power.

Moreover, the 'Baseline' column in Table 10 presents the implied changes in other quantities that arise from these gains. Stemming from higher TFP and higher production labor, output is 14.6% higher under the efficient allocation despite the fact that the number of firms is 11.3% lower. This higher production with fewer firms is made possible by the fact that concentration of sales among the top 5% largest firms is 39.2% larger than in the equilibrium. In addition to the calibrated model, Table 10 also presents similar results for two counterfactual values of ϕ. In the remainder of this section, we dive into dissecting these changes and study the underlying forces that shape these gains.

Implications for Misallocation It is known that without endogenous customer acquisition, this model would predict quantitatively small losses from misallocation (see Edmond

Table 10: Comparison with Efficient Allocation

	Endogenous $m_{i,t}$		
	$\phi = 0.25$	Baseline	$\phi = 0.75$
TFP	24.1	10.8	3.2
Output	27.5	14.6	7.7
Number of firms	-41.9	-11.3	-2.6
Employment	-5.0	2.1	4.4
Production	5.3	6.0	7.0
Welfare	37.9	13.6	4.0
Agg. markup	-27.8	-22.8	-19.1
Top 5% sales share	88.8	39.2	15.5

Notes: The table compares aggregate variables between the social planner's allocation and the equilibrium allocation. Differences are reported as percent deviations from equilibrium allocations. Three comparisons are presented by varying the value of ϕ, while keeping the remaining parameters fixed at the values in the baseline calibration.

et al., 2018, who find losses from misallocation to be in a range of 0.8% to 1.8%). However, by explicitly modeling endogenous customer acquisition, in a way that is consistent with our motivating facts, we find these losses to be significantly higher at 10.8%.

To further analyze the increase in aggregate productivity, we consider the decomposition of TFP derived in Baqaee and Farhi (2019) and separate the *allocative efficiency gains* from *technological change*. For us, allocative efficiency refers to how differently the planner allocates resources across firms, while technological change is a manifestation of the different entry and exit policies that the planner adopts. Formally, let $Z(N_t, \mathcal{A}_t)$ denote the aggregate productivity implied by the set of operating firms N_t with an allocation rule $\mathcal{A}_t \equiv (l_{i,p,t})_{i \in N_t}$ among them. Then, we can decompose the difference in TFPs across two allocations as

$$\underbrace{\ln\left(\frac{Z(N_t^*, \mathcal{A}_t^*)}{Z(N_t, \mathcal{A}_t)}\right)}_{\Delta \text{ TFP} = 10.8\%} = \underbrace{\ln\left(\frac{Z(N_t, \mathcal{A}_t^*)}{Z(N_t, \mathcal{A}_t)}\right)}_{\Delta \text{ Allocative Efficiency} = 7.8\%} + \underbrace{\ln\left(\frac{Z(N_t^*, \mathcal{A}_t^*)}{Z(N_t, \mathcal{A}_t^*)}\right)}_{\Delta \text{ Entry/Exit Efficiency} = 3.0\%} . \tag{6.1}$$

The first term on the right hand side of Equation (6.1) shows that, keeping the set of operating firms fixed, almost 75% of the efficiency gains under the planner's allocations are due to allocative efficiency gains. This is in fact the most important consequence of endogenous customer acquisition: having the ability to reallocate customers across firms, the planner shifts the distribution of customers towards the top of the productivity distribution, and hence is able to allocate higher amounts of production labor towards them.

The extensive margin of demand is the key to the difference between our higher TFP

gains relative to Edmond et al. (2018): once the extensive margin of demand is shut down, the planner in their model can achieve higher aggregate productivity only by shifting demand on the intensive margin towards more productive firms. However, since varieties are weak substitutes, distorting the distribution of relative demand is costly. These costs are even higher when demand is more elastic at higher quantities (as with Kimball preferences or any semi-kinked demand system), which is why the aggregate efficiency gains in Edmond et al. (2018) are small.

In our model, however, the planner has an extra choice, which is the distribution of customers across firms. As a result, she does not face a trade-off as in conventional models and shifts demand only on the extensive margin. Completely equalizing relative consumption across individuals ($q^*_{i,t} = 1$), the efficient allocation achieves much higher aggregate productivity by simply allocating *more customers* towards more productive firms ($m^*_{i,t} \propto z_{i,t}^{\frac{1}{1-\alpha}}$)—as seen in Figure F.8, which shows a comparison of the allocation of customers between the equilibrium and the efficient allocation. As a consequence of this reallocation of customers, concentration of sales among the top 5% of firms increases by 39.2%. Higher concentration of customers across more productive firms is efficient to the point that marginal costs of production are equalized across all firms.

It is important to note that the efficient allocation of customers across firms is not restricted by the decreasing returns to scale in advertising, even though that the planner is subject to the same advertising technology as in the equilibrium. This follows from the fact that the planner internalizes the business-stealing externalities of advertising, and by Lemma 2 can implement any desired distribution of customers.

Finally, while the optimal allocation of resources accounts for around 75% of the change in aggregate TFP, the remaining 25% is explained by sheer compositional changes in the distribution of productivity, i.e., technological change. Under the efficient allocation, the planner is more selective in allowing firms to enter and ends up choosing a higher productivity cutoff for entry and exit of firms. A more selective policy increases productivity because it increases the average productivity of firms that operate in the economy, and it implies a fewer number of operating firms—since the planner does not control the measure of potential entrants that are born in every period. This last observation brings us to the next substantial difference between our model and the conventional models.

The Optimal Number of Firms We start by reviewing the usual costs and benefits of having more firms in the economy and then discuss the new mechanism that comes into play in our model. In conventional models, the optimal number of firms are affected by the interaction of three forces: decreasing returns to scale, love for variety and aggregate overhead costs. On one side, with decreasing returns to scale at the firm level, having more

firms increases the aggregate efficiency by dividing resources across a larger number of production units. Moreover, with love for variety, even fixing the average output produced by a larger set of firms, the household enjoys the resultant *aggregated* output more and hence the economy experience a higher productivity.[28] While these forces form the benefits of a higher number of firms, the cost is usually modeled either as a fixed entry cost for every firm or, as we model here, a stream of overhead costs over time, both of which lead to an optimal finite number firms in the equilibrium.

Our model shares all these forces with the conventional models but has an additional force which is the allocation of customers across firms. In conventional models, to increase productivity the planner would have to bring in more firms through the bottom of the productivity distribution by lowering the entry productivity cutoff; in our model, on the other hand, the planner can achieve the same objective by allocating customers towards firms at the top of the productivity distribution. However, since the number of customers is fixed, higher concentration at the top comes at the cost of fewer customers at the bottom of the distribution of productivity, which in turn reduces the social value of such firms (as shown in Proposition 4). Hence, with this additional instrument, our planner is able to achieve a higher productivity without having to pay for the overhead costs of more firms, which leads to a lower number of firms in the efficient allocation and increases the welfare of the household by 1.6% as shown in Equation (6).

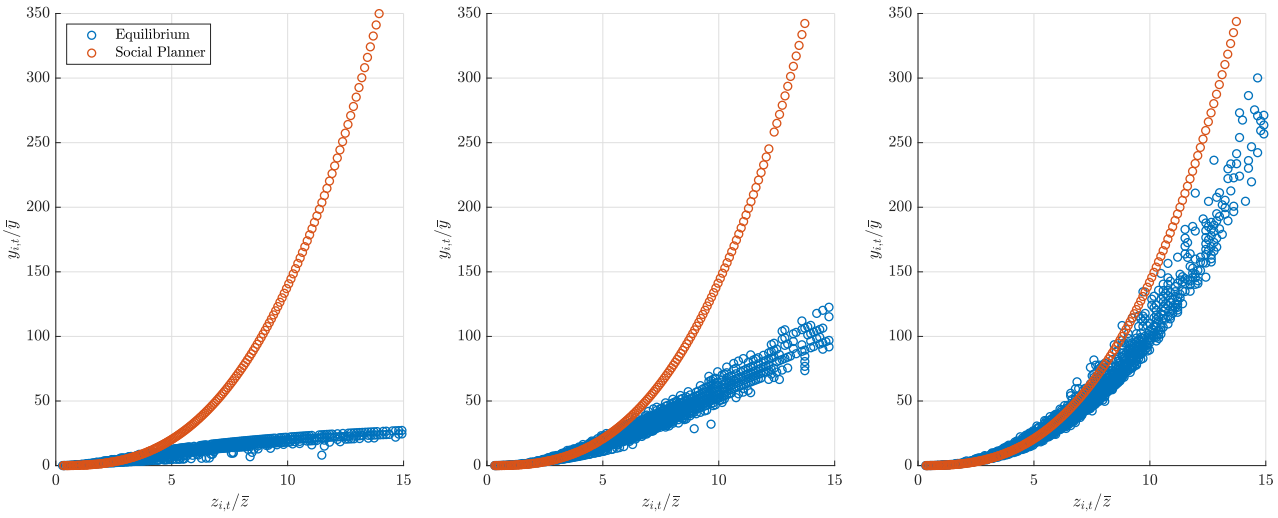
To summarize, more productive firms acquire fewer customers in the equilibrium than is efficient because of the decreasing returns to advertising, which makes customers "cheap" to acquire for less productive firms and leads to excessive entry in the equilibrium.

Aggregate Labor Supply Two forces work in opposite directions in affecting the differences of aggregate labor supply between the efficient and equilibrium allocations. On one hand, the more selective policy of the planner for entry and exit reduces the amount of labor required for financing the overhead costs of operating firms. On the other hand, production labor is underutilized in the equilibrium due to the aggregate market power of firms. The 'Baseline' column of Table 10 shows that while labor allocated towards production goes up by 6% under the efficient allocation, which together with the higher aggregate TFP contributes towards the 14.6% increase in output, the aggregate labor goes up only by 2.1%, as it is mitigated by the lower use of labor in financing the entry cost of firms.

[28] Both of these forces can be summarized by the following simple example inspired by Edmond et al. (2018): consider an economy with N firms indexed by i, where every firm produces with $y_i = l_i^{\alpha}$ and aggregate output is given by a CES aggregator, $Y = [\int_0^N y_i^{\theta-1} di]^{\theta}$. For given amount of aggregate labor, L, every firms gets to produce $y_i = (L/N)^{\alpha}$ and the aggregate output is given by $Y = N^{\theta-\alpha} L^{\alpha}$. Now if we shut down love for variety ($\theta = 1$), productivity is $N^{1-\alpha}$ which increases with N. If we shut down decreasing returns to scale ($\alpha = 1$), productivity is $N^{\theta-1}$ indicating higher productivity due to love for variety with larger N.

The Role of Returns to Scale in Marketing While, for the planner, the only relevant margin in allocating customers is returns to scale in production, it is important to note that the efficiency gains from reallocation of customers depend on the returns to scale for customer acquisition, ϕ. A larger returns to scale in customer acquisition would imply that more productive firms would invest more in customer acquisition, which is desirable from the perspective of the efficient allocation. Figure 7 shows the scatter plot of firms' productivity and output for both the equilibrium and social planner's allocation and for three different values of ϕ (a low value, the calibrated value and a high value). The figure shows that with a larger ϕ the equilibrium allocation of customers is closer to that of the planner.

Figure 7: Allocation of Output: Equilibrium vs. Efficient Allocation



(a) Low ϕ (b) Baseline (c) High ϕ

Notes: This figure shows a scatter plot between relative productivity $z_{i,t}/\bar{z}$ and relative output $y_{i,t}/\bar{y}$, for both the equilibrium and the social planner's allocation. We present three plots by varying the value of ϕ, while keeping the remaining parameters fixed at the values in the baseline calibration. Low ϕ corresponds to 0.25, baseline to 0.53, and high to 0.75.

Moreover, the $\phi = 0.25$ and $\phi = 0.75$ columns of Table 10 show how the allocation of customers is solely responsible for the large efficiency gains under the planner's allocation. By simply allowing ϕ to be larger, the equilibrium welfare losses drop from 38% in the case of $\phi = 0.25$ to only 4% with $\phi = 0.75$. When ϕ is larger, in the equilibrium, more productive firms grow mainly through acquiring more customers (higher m) rather than selling more per customer (higher q). As a result, they produce for more customers but sell less per customer, which also implies that they charge lower markups. Hence, aggregate TFP, output and concentration increase, but aggregate markups decrease and the economy gets closer to the efficient allocation.

7 Conclusion

In this paper, we revisit the role of the extensive and intensive margins of demand in firms' market share and market power. Using a dataset that merges information from the consumer and the producer sides, we document that while firms' sales grow mainly through acquiring more customers, their market power is only correlated with their average sales per customer. Moreover, we find that firms' non-production costs are associated with their customer acquisition but not customer retention or sales per customer.

Guided by these empirical findings, we develop and quantify a model that micro-founds the relationship between market power and concentration in the extensive and intensive margins. In our model, while firms hold market power over each customer, the total number of customers acts as a demand shifter. The model provides a new perspective on the relationship between firm size and market power. Firms that are big due to a larger customer base, have lower market power relative to equally big firms with higher sales per customer. Our model predicts higher concentration than conventional models, but lower aggregate market power. Nonetheless, we find substantive welfare gains under the efficient allocation that stems from the new Pareto frontier of the economy under endogenous customer acquisition.

Our analysis sheds light on the effectiveness of policies that target concentration, profits and market power. In particular, our model highlights a new unintended consequence from policies that target only firms' market power. In our model, although market power is distortionary, it compensates more productive firms for their investment in customer acquisition and improves the allocation of customers. Thus, policies that target larger firms disproportionately may have adverse effects through the misallocation of customers. If more productive firms are taxed for their larger sales due to larger customer bases, on the margin they will sell to fewer customers at lower prices but higher markups—both of which are inefficient from a social perspective.

References

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

Andrews, Isaiah, Matthew Gentzkow, and Jesse M Shapiro, "Measuring the Sensitivity of Parameter Estimates to Estimation Moments," *The Quarterly Journal of Economics*, 2017, *132* (4), 1553–1592.

Argente, David, Munseob Lee, and Sara Moreira, "Innovation and Product Reallocation in the Great Recession," *Journal of Monetary Economics*, 2018, *93*, 1–20.

_ , _ , and _ , "The Life Cycle of Products: Evidence and Implications," *Available at SSRN 3163195*, 2019.

Arkolakis, Costas, "Market Penetration Costs and the New Consumers Margin in International Trade," *Journal of Political Economy*, December 2010, *118* (6), 1151–1199.

Arnoud, Antoine, Fatih Guvenen, and Tatjana Kleineberg, "Benchmarking Global Optimizers," Technical Report, National Bureau of Economic Research 2019.

Asker, John, Allan Collard-Wexler, and Jan De Loecker, "Dynamic Inputs and Resource (Mis)allocation," *Journal of Political Economy*, January 2014, *122* (5), 1013–1063.

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

Autor, David, David Dorn, Lawrence F Katz, Christina Patterson, and John Van Reenen, "The Fall of the Labor Share and the Rise of Superstar Firms," *The Quarterly Journal of Economics*, 02 2020.

Bagwell, Kyle, "The economic analysis of advertising," *Handbook of industrial organization*, 2007, *3*, 1701–1844.

Baqaee, David Rezza and Emmanuel Farhi, "Productivity and Misallocation in General Equilibrium*," *The Quarterly Journal of Economics*, September 2019, *135* (1), 105–163.

Basu, Susanto, "Comment On:" Implications of State-Dependent Pricing for Dynamic MacRoeconomic Modeling"," *Journal of Monetary Economics*, 2005, *52* (1), 243–247.

Bigio, Saki and Jennifer La'O, "Distortions in Production Networks*," *The Quarterly Journal of Economics*, May 2020, *135* (4), 2187–2253.

Bils, Mark, "Pricing in a Customer Market," *The Quarterly Journal of Economics*, November 1989, *104* (4), 699–718.

Bolton, Ruth N, P K Kannan, and Matthew D Bramlett, "Implications of Loyalty Program Membership and Service Experiences for Customer Retention and value," *Journal of the Academy of Marketing Science*, 2000, *28* (1), 95–108.

Bond, Steve, Arshia Hashemi, Greg Kaplan, and Piotr Zoch, "Some Unpleasant Markup Arithmetic: Production Function Elasticities and their Estimation from Production Data," Technical Report, National Bureau of Economic Research 2020.

Bornstein, Gideon, "Entry and Profits in an Aging Economy: The Role of Consumer Inertia," Technical Report 2018. Mimeo.

Buera, Francisco J, Joseph P Kaboski, and Yongseok Shin, "Finance and Development: A Tale of Two Sectors," *American Economic Review*, 2011, *101* (5), 1964–2002.

Burstein, Ariel, Vasco M Carvalho, and Basile Grassi, "Bottom-up Markup Fluctuations," Working Paper 27958, National Bureau of Economic Research October 2020.

Cabral, Luís, "Dynamic Pricing in Customer Markets With Switching Costs," *Review of Economic Dynamics*, April 2016, *20* (C), 43–62.

Clementi, Gian Luca and Berardino Palazzo, "Entry, Exit, Firm Dynamics, and Aggregate Fluctuations," *American Economic Journal: Macroeconomics*, July 2016, *8* (3), 1–41.

Covarrubias, Matias, Germán Gutiérrez, and Thomas Philippon, "From Good to Bad Concentration? US Industries Over the Past 30 Years," *NBER Macroeconomics Annual*, January 2020, *34*, 1–46.

Crouzet, Nicolas and Janice C Eberly, "Understanding Weak Capital Investment: The Role of Market Concentration and Intangibles," *National Bureau of Economic Research Working Paper Series*, May 2019.

David, Joel M, Hugo A Hopenhayn, and Venky Venkateswaran, "Information, Misallocation, and Aggregate Productivity," *The Quarterly Journal of Economics*, 2016, *131* (2), 943–1005.

Davis, Steven J and John Haltiwanger, "Gross Job Creation, Gross Job Destruction, and Employment Reallocation," *The Quarterly Journal of Economics*, 1992, *107* (3), 819–863.

_ , John C Haltiwanger, Scott Schuh et al., "Job Creation and Destruction," *MIT Press Books*, 1998, *1*.

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

Dinlersoz, Emin M and Mehmet Yorukoglu, "Information and Industry Dynamics," *American Economic Review*, 2012, *102* (2), 884–913.

Dotsey, Michael and Robert G King, "Implications of State-Dependent Pricing for Dynamic MacRoeconomic Models," *Journal of Monetary Economics*, 2005, *52* (1), 213–242.

Drozd, Lukasz A and Jaromir B Nosal, "Understanding International Prices: Customers as Capital," *American Economic Review*, February 2012, *102* (1), 364–395.

Edmond, Chris, Virgiliu Midrigan, and Daniel Yi Xu, "How Costly Are Markups?," Technical Report, National Bureau of Economic Research 2018.

Einav, Liran, Peter J Klenow, Jonathan D Levin, and Raviv Murciano-Goroff, "Customers and Retail Growth," 2020.

Elsby, Michael WL and Ryan Michaels, "Marginal Jobs, Heterogeneous Firms, and Unemployment Flows," *American Economic Journal: Macroeconomics*, 2013, *5* (1), 1–48.

Fitzgerald, D and A Priolo, "How Do Firms Build Market Share?," 2018.

Fitzgerald, Doireann, Stefanie Haller, and Yaniv Yedid-Levi, "How Exporters Grow," *National Bureau of Economic Research Working Paper Series*, January 2016.

Foster, Lucia, John Haltiwanger, and Chad Syverson, "The Slow Growth of New Plants: Learning About Demand?," *Economica*, December 2015, *83* (329), 91–129.

Garthwaite, Craig L, "Demand spillovers, combative advertising, and celebrity endorsements," *American Economic Journal: Applied Economics*, 2014, *6* (2), 76–104.

Gilchrist, Simon, Raphael Schoenle, Jae Sim, and Egon Zakrajšek, "Inflation Dynamics During the Financial Crisis," *American Economic Review*, 2017, *107* (3), 785–823.

Gopinath, Gita and Oleg Itskhoki, "Frequency of Price Adjustment and Pass-Through," *The Quarterly Journal of Economics*, 2010, *125* (2), 675–727.

_ , Pierre-Olivier Gourinchas, Chang-Tai Hsieh, and Nicholas Li, "International Prices, Costs and Mark-up differences," *American Economic Review*, 2011, *101* (6), 2450–86. Previously circulated under the title "Estimating the Border Effect: Some New Evidence".

Gourio, Francois and Leena Rudanko, "Customer Capital," *Review of Economic Studies*, 2014, *81* (3), 1102–1136.

Haltiwanger, John, Ron S Jarmin, and Javier Miranda, "Who Creates Jobs? Small Versus Large Versus Young," *Review of Economics and Statistics*, 2013, *95* (2), 347–361.

Hartmann, Wesley R and Daniel Klapper, "Super bowl ads," *Marketing Science*, 2018, *37* (1), 78–96.

Hong, Sungki, "Customer Capital, Markup Cyclicality, and Amplification," Technical Report, Federal Reserve Bank of St. Louis, St. Louis, MO, USA 2017.

Hopenhayn, Hugo A, "Entry, Exit, and Firm Dynamics in Long Run Equilibrium," *Econometrica*, 1992, pp. 1127–1150.

Hopenhayn, Hugo, Julian Neira, and Rish Singhania, "The Rise and Fall of Labor Force Growth: Implications for Firm Demographics and Aggregate Trends," 2018. Mimeo.

Hottman, Colin J, Stephen J Redding, and David E Weinstein, "Quantifying the Sources of Firm Heterogeneity," *The Quarterly Journal of Economics*, 2016, *131* (3), 1291–1364.

Hsieh, Chang-Tai and Peter J Klenow, "Misallocation and Manufacturing TFP in China and India," *Quarterly Journal of Economics*, November 2009, *124* (4), 1403–1448.

Kaplan, Greg and Piotr Zoch, "Markups, Labor Market Inequality and the Nature of Work," *National Bureau of Economic Research Working Paper Series*, February 2020.

Kimball, Miles, "The Quantitative Analytics of the Basic Neomonetarist Model," *Journal of Money, Credit and Banking*, 1995, *27* (4), 1241–77.

Klenow, Peter J and Jonathan L Willis, "Real Rigidities and Nominal Price Changes," *Economica*, 2016, *83* (331), 443–472.

Lee, Yoonsoo and Toshihiko Mukoyama, "Productivity and Employment Dynamics of US Manufacturing Plants," *Economics Letters*, 2015, *136*, 190–193.

Midrigan, Virgiliu and Daniel Yi Xu, "Finance and Misallocation: Evidence From Plant-Level Data," *American Economic Review*, 2014, *104* (2), 422–58.

Mittal, Vikas and Wagner A Kamakura, "Satisfaction, Repurchase Intent, and Repurchase Behavior: Investigating the Moderating Effect of customer characteristics," *Journal of marketing research*, 2001, *38* (1), 131–142.

Nakamura, Emi and Dawit Zerom, "Accounting for Incomplete Pass-Through," *The Review of Economic Studies*, 2010, *77* (3), 1192–1230.

_ **and Jón Steinsson**, "Price Setting in Forward-Looking Customer Markets," *Journal of Monetary Economics*, April 2011, *58* (3), 220–233.

Neiman, Brent and Joseph S Vavra, "The Rise of Niche Consumption," Technical Report, National Bureau of Economic Research 2019.

Ottonello, Pablo and Thomas Winberry, "Financial Heterogeneity and the Investment Channel of Monetary Policy," Technical Report, National Bureau of Economic Research 2018.

Paciello, Luigi, Andrea Pozzi, and Nicholas Trachter, "Price Dynamics With Customer Markets," *International Economic Review*, October 2018, *60* (1), 413–446.

Perla, J, "A Model of Product Awareness and Industry Life Cycles," 2019.

Peters, Michael, "Heterogeneous markups, growth, and endogenous misallocation," *Econometrica*, 2020, *88* (5), 2037–2073.

Phelps, Edmund S and Sidney G Winter, "Optimal Price Policy Under Atomistic Competition," *Microeconomic foundations of employment and inflation theory*, 1970, pp. 309–337.

Ravn, Morten, Stephanie Schmitt-Grohé, and Martin Uribe, "Deep Habits," *The Review of Economic Studies*, January 2006, *73* (1), 195–218.

Restuccia, Diego and Richard Rogerson, "Policy Distortions and Aggregate Productivity With Heterogeneous Establishments," *Review of Economic Dynamics*, October 2008, *11* (4), 707–720.

Rotemberg, Julio J and Michael Woodford, "Oligopolistic Pricing and the Effects of Aggregate Demand on Economic Activity," *Journal of Political Economy*, 1992, *100* (6), 1153–1207.

_ **and** _ , "The Cyclical Behavior of Prices and Costs," *Handbook of macroeconomics*, 1999, *1*, 1051–1135.

Sedláček, Petr and Vincent Sterk, "The Growth Potential of Startups Over the Business Cycle," *American Economic Review*, October 2017, *107* (10), 3182–3210.

Sinkinson, Michael and Amanda Starc, "Ask your doctor? Direct-to-consumer advertising of pharmaceuticals," *The Review of Economic Studies*, 2019, *86* (2), 836–881.

Stroebel, Johannes and Joseph Vavra, "House Prices, Local Demand, and Retail Prices," *Journal of Political Economy*, 2019, *127* (3), 1391–1436.

Syverson, C, "Macroeconomics and Market Power: Facts, Potential Explanations and Open Questions, Brookings Economic Studies," *Brookings Institution, Washington DC*, 2019.

Traina, James, "Is Aggregate Market Power Increasing? Production Trends Using Financial Statements," 2019. Mimeo.

Wasi, Nada and Aaron Flaaen, "Record Linkage Using Stata: Preprocessing, Linking, and Reviewing Utilities," *The Stata Journal*, 2015, *15* (3), 672–697.

Young, Eric R, "Solving the Incomplete Markets Model With Aggregate Uncertainty Using the Krusell–Smith algorithm and non-stochastic simulations," *Journal of Economic Dynamics and Control*, 2010, *34* (1), 36–41.

APPENDIX FOR ONLINE PUBLICATION

A Further data description

This section further describes the process to clean the Nielsen and Compustat data used in this paper. For this, we follow the previous studies that use these two datasets (Hottman et al. 2016 for Nielsen and De Loecker et al. 2020; Traina 2019 for Compustat).

A.1 Nielsen and GS1: Variables and Data Cleaning

We construct the following variables from the Nielsen Homescan Panel, Promodata, and GS1 company data:

- *UPC*: scanner-level product identifier available in Nielsen Homescan Panel and Promo data.
- *GS1 Company Identifier*: GS1 company id. GS1 provides both UPC and firm identifiers along with company name and headquarter location. This information allows us to identify firm boundaries in the Nielsen data and further merge the data with Compustat.
- *Sales*: We define sales as the sum of the total expenditures of households at different levels of aggregation: UPC-year, firm-year, and group-firm-year. We use sample weights (projection factor) to make the Nielsen household sample representative at the national level.
- *Number of Customers*: For each firm, product (UPC), or group-firm, we aggregate the number of households with the sample weight adjustment.
- *Sales per Customer*: At each level of aggregation, we define average sales per customer as total sales divided by the number of customers.

Sample Selection of Homescan Panel data The Homescan Panel data we use covers 2004-2016.

1. Following Neiman and Vavra (2019), we balance the product modules across years to exclude the effect of entry and exit of modules, which mainly arises from name changes and potentially adds measurement error.

2. To make the sample representative, we drop "magnet products" in Nielsen data which are fresh produce and other items without barcodes.

3. We drop products that do not have a product group identifier to make the analysis consistent across different specifications.

4. There is a small number of observations for which the sampling year of the household is different than the year when their purchases took place. They reflect the fact that households were sampled in late December. While the corresponding purchases are

recorded as the current year, their household panel years are recorded as the following year. We drop these observations to use coherent sample weights across households and years.

We restrict the sample in the Promodata to the years 2006-2011 since the data is incomplete for the other years. Following this data's manual, we use both active and inactive files and drop the duplicated observations in inactive files. We adjust for multi-package and unit size in utilizing the UPC-level information; we do the same when we merge in price information from Homescan Panel data. We exclude the small number of markets that are not common across the Homescan Panel and Promodata.

A.2 Compustat: Variables and Data Cleaning

We download and construct the following variables from Compustat:

- *Global company key* (mnemonic `gvkey`): Compustat's firm id.
- *Year* (mnemonic `fyear`): the fiscal year.
- *Selling, general and administrative expense* (mnemonic `XSGA`): the SG&A sums "all commercial expenses of operation (such as, expenses not directly related to product production) incurred in the regular course of business pertaining to the securing of operating income." They include expenses such as marketing and advertising expenses, research and development, accounting expenses, delivery expenses, etc.
- *Costs of goods sold* (mnemonic `COGS`): the COGS sums all "expenses that are directly related to the cost of merchandise purchased or the cost of goods manufactured that are withdrawn from finished goods inventory and sold to customers." They include expenses such as labor and related expenses (including salary, pension, retirement, profit sharing, provision for bonus and stock options, and other employee benefits), operating expense, lease, rent, and loyalty expense, write-downs of oil and gas properties, and distributional and editorial expenses.
- *Operating expenses, total* (mnemonic `XOPR`): OPEX represents the sum of COGS, SG&A and other operating expenses.
- *Sales (net)* (mnemonic `SALE`): this variable represents gross sales, for which "cash discounts, trade discounts, and returned sales and allowances for which credit is given to customer" are discounted from the final value.
- *Capital*: we calculate capital in two ways. First, we simply set capital to be equal to the gross property, plant and equipment value (mnemonic `PPEGT`) deflated by the investment goods deflator from NIPA's nonresidential fixed investment good deflator (line 9). For our second measurement of capital, we use the perpetual inventory method—we set the first observation of each firm to be equal to the gross property, plant and equipment value and for susbsequent years we add the difference from $netPPE_t$ (mnemonic `PPENT`) and $netPPE_{t-1}$. We also deflate the difference in `PPENT` by the investment goods deflator from NIPA's nonresidential fixed investment good deflator.
- *Company's initial public offering date* (mnemonic `IPODATE`).

- *Age*: given that the initial public offering date was missing for a large portion of our dataset, we calculated age as the fiscal year of a given observation minus the first year that we observe a firm in the dataset. According to Compustat, a firm enters the dataset after it starts providing consistent accessible annual reports trading on a U.S. exchange market, i.e. after its IPO. Following Haltiwanger et al. (2013), we exclude the first fifteen years of the dataset for all analyses using age and we group together firms older than sixteen years, because we do not know for certain a firm's IPO date for firms that were in the Compustat data since the first year.

We used NIPA Table 1.1.9. GDP deflator (line 1) to generate the real value for the variables `sale`, `COGS`, `XOPR`, and `XSGA`.

Sample Selection We downloaded the dataset "Compustat Annual Updates: Fundamentals Annual," from Wharton Research Data Services, from Jan 1950 to Dec 2016. The following options were chosen:

- Consolidated level: C (consolidated)
- Industry format: INDL (industrial)
- Data format: STD (standardized)
- Population source: D (domestic)
- Currency: USD
- Company status: active and inactive

We took the following steps for the cleaning process:

1. To select American companies, we filtered the dataset for companies with Foreign Incorporation Code (FIC) equal to "USA."

2. We replace industry variables (`sic` and `naics`) by their historical values whenever the historical value is not missing.

3. We drop utilities (`sic` value in the range [4900, 4999]) because their prices are very regulated and financials (`sic` value in the range [6000,6999]) because their balance sheets are exceptionally different than the other firms in the analysis.

4. To ensure quality of the data, we drop missing or non-positive observations for sales, COGS, OPEX, sic 2-digit code, gross PPE, net PPE, and assets. We also exclude observations in which acquisitions are more than 5% of the total assets of a firm.

5. A portion of the data missing for sales, COGS, OPEX, and capital in between years for firms. We input these values using a linear interpolation, but we do not interpolate for gaps longer than two years. This exercise inputs data for 4.6% of our sample.

A.3 The Coverage of the Nielsen-Compustat Sample

There are approximately 300 firms identified in Compustat that can be matched with the Nielsen data in 2004-2016. Although the number of firms we matched is small, they account for a significant fraction of total sales, number of UPCs, and observations in the Nielsen Homescan Panel data, as shown in Table A.1.

	Sales (b)	# of UPCs (k)	# of Obs. (m)
Nielsen-Compustat Sample	94.5	114.9	12.1
Nielsen Sample	421.2	698.9	51.6
Share (%)	22.4	16.4	23.5

Table A.1: The Coverage of the Nielsen-Compustat Sample

Note: Sales is the projection-factor weighted sales in Nielsen data and is denoted in billions US dollars. # of UPCs is in thousand UPCs, and # of Obs. is in millions of observations. All variables are annual averages.

B Additional Empirical Results

B.1 Robustness

This section provides additional results to confirm the robustness of our main empirical findings.

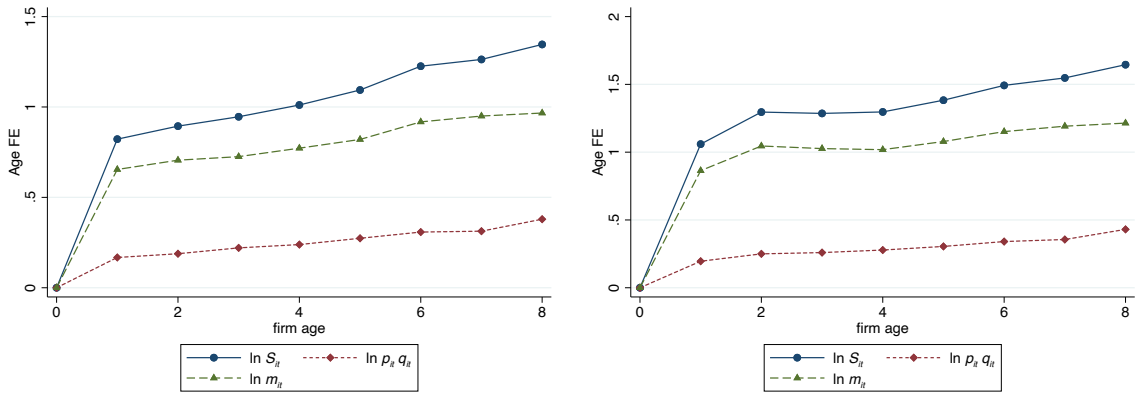
B.1.1 Firm Sales Growth Decomposition

One concern in Figure 1 is that some firms might only appear temporarily in our data not because of their actual behavior but due to the sampling error. For example, it could be that households in our sample do not happen to purchase a firm's product even though the product was purchased outside of the sample. In this case, the average value of sales, number of customers, and sales per customer of young firms in our analyses might be confounded with those of old firms.

To address the concern of the sampling error, Figure B.1 uses only those firms that appear at least three or five consecutive years. The results still show that the number of customers is a primary factor that generates an increase in firms' sales over time. There is a steeper increase in sales in the firm's early-stage than our baseline results in Figure 1. The results are intuitive since firms that survive for several years are likely to generate more sales at the beginning relative to firms that could not survive. Overall, the robustness results suggest that the sampling errors are not the first-order concerns in our analyses.

Another concern is that firms might sell their products in a different number of months over different years. For example, some firms might enter in late November or December but sell their products over many months of the subsequent years. To adjust the differences, we calculate the average monthly sales over a year per firm and redo the decomposition exercise in Figure B.2. There is a smaller increase in firms' sales at age 1, suggesting that some firms enter the late month of the initial year. The relative importance of the number of

Figure B.1: Decomposition of Firm Sales Growth by Firm Age: Survivors



(a) Survive 3 Consecutive Years (b) Survive 5 Consecutive Years

Notes: Figures B.1a and Figure B.1b replicate Figure 1 by using the firms that appear at least 3 and 5 consecutive years, respectively. There are 32,242 number of observations and 6,400 firms used in Figures B.1a and 19,603 number of observations and 2,997 firm used in Figure B.1b.

customers in explaining sales remain the same, explaining approximately 70% of sales on average.

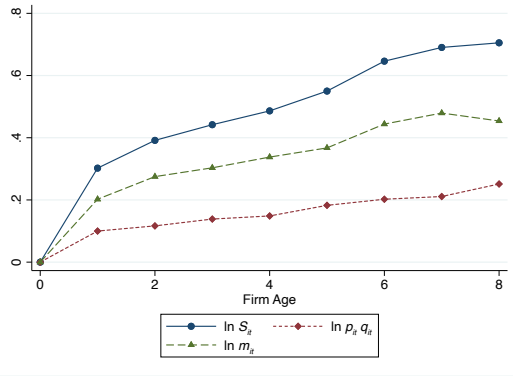


Figure B.2: Decomposition of Firm Sales Growth by Firm Age: Monthly Sales

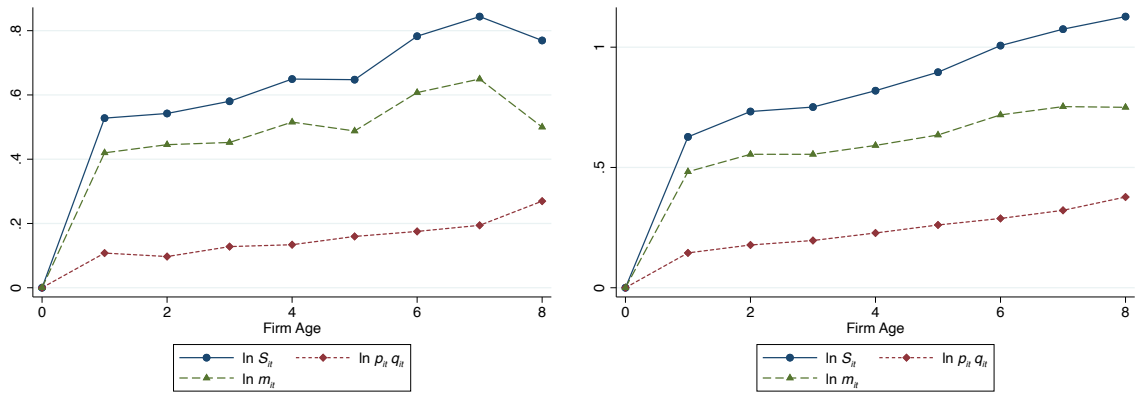
Notes: Figure B.2 replicates Figure 1 by using average monthly sales per firm and year instead of yearly sales.

Also, one might be worried that the empirical pattern we observe might not apply to the products outside of our sample, which is restricted to products that have barcode. For example, it could be that for more durable products that customers purchase occasionally, firms might not be able to grow through the number of customers because they may not face customers every year.

Given that there is no other consumer-producer matched data (to the best of our knowledge), we utilize our data—which covers a substantial fraction of consumer goods with a wide variety of products—to understand the underlying differences between durable and non-durable products. We carefully follow Argente et al. (2019) and define the product group-level durability by using the information on the number of shopping trips. We count the average yearly number of customer trips to purchase products in each product

group and divide product groups into durable and non-durable products based on the median value. The durable products include "LIGHT BULBS, ELECTRIC GOODS", "HARDWARE, TOOLS", and "AUTOMOTIVE", and the non-durable products include "MILK", "SNACKS", and "BEER".

Figure B.3: Decomposition of Firm Sales Growth by Firm Age: by Durability



(a) $>$ p50 durability (b) $\leq$ p50 durability

Notes: Figures B.3a and Figure B.3b replicate Figure 1 by dividing firms based on the durability of products they sell. There are 19,988 observations and 5,050 firms used in Figures B.3a and 29,816 observations and 7,323 firms used in Figure B.3b.

Figure B.3 presents the results. Regardless of analyzing durable or non-durable products, firms mainly grow through the number of customers. The importance of the number of customers remains by redefining durable goods based on the 75th percentile or by analyzing the variance decomposition of either durable or non-durable products.

B.1.2 Markups, Sales per customers, and Sales

Table B.1 replicate Table 3 by replacing the number of customers with the sales. We replace the number of customers with sales so that our independent variables have the same unit. Regardless of controlling sales or not, our results still show a strong correlation between markups and sales per customer, suggesting the importance of the sales per customer in understanding the firm-level markups.

B.1.3 Alternative Measures of Markups

This section revisits our analysis of the relationship between markups and the size of different margins of firm's demand by considering two alternative approaches to measure price-cost markup and account for marginal costs. Our baseline analysis focuses on firm-level markups and public firms. Here, we should that we find similar results when expanding the scope to product-level markups and analyzing a much broader set of firms in the Nielsen data.

In our first alternative approach, we use the difference between the retail-level price and the wholesale cost at the product level as a measure of markups. For this, we use micro-data on the UPC-market-year-level wholesale cost from the Nielsen PromoData available

Table B.1: Markups, Sales, and Sales per Customer

	(1)	(2)	(3)	(4)	(5)
ln $p_{it}q_{it}$	0.094***	0.093***	0.058**	0.056**	0.057**
	(0.031)	(0.032)	(0.023)	(0.023)	(0.025)
ln S_{it}	-0.002	-0.002	0.002	0.002	0.003
	(0.006)	(0.006)	(0.007)	(0.007)	(0.007)
Observations	2433	2433	2433	2433	2433
R^2	0.046	0.047	0.311	0.313	0.338
Year FE		✓		✓	
SIC FE			✓	✓	
SIC-year FE					✓

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$; standard errors are clustered at the firm-level. The markups are measured as sales-to-COGS ratio. SIC is a two-digit SIC code, and all Nielsen variables are projection-factor adjusted.

through the Chicago Booth Kilts Center. The PromoData records in wholesale costs by UPC, market, and year and is collected from 12 national wholesalers in the period 2006-2012 (we restrict our sample to the years 2006-2011 since there is a substantial number of missing observations in 2012). In total, we use data from 45 markets (examples of markets are Chicago, Los Angeles, and Atlanta). Given that the data lacks sales information, we take a simple geometric average of wholesale costs by UPC, market, and year after adjusting for the package size. We then combine this information with the retail-level price, sales, and sales per household from the Nielsen Homescan Panel data and measure the retail-level markup at the product (UPC) level as the difference between the retail price and the wholesale cost (we drop negative values of the measured markups). A similar approach has been followed by Gopinath et al. (2011) and Stroebel and Vavra (2019).

The previous measure of markups assumes that other costs, such as wages and capital expenditures, do not confound the observed relationship among markups, average sales per customer, and the number of customers. However, presumably, there exist sources of variation in marginal costs across retailers as well, in addition to the variation arising from differences in product-level wholesale costs. To alleviate these concerns, we exploit the rich variation in the data and control for such variation in marginal costs by incorporating in the regression various sets of fixed effects, as the approach followed by Fitzgerald et al. (2016). Our measure of markup varies at the UPC, year, market, and retailer level, so we progressively include different combinations of fixed effects at those levels to absorb differences in marginal costs that could potentially confound the relationship. For example, the fact that retailers sell the same UPC in multiple locations allows us to control for marginal costs at the retailer-year-UPC level. The underlying assumption is that the retailer's marginal cost of selling a given UPC is common across markets. Similarly, we can account for differential distribution costs by incorporating a set of UPC-market-year fixed effects. The final sample includes 7,802 UPCs, 45 markets, 6 years, and 167 retailers.

Table B.2 confirms the results about the relationship between markups, average sales per

customer, and the number of customers reported in Table 3. In the first two columns, we measure markups following the first alternative approach, while in the remaining columns we present the results for the second approach. Column (1) shows that markups are strongly positively correlated with sales per customer. The relationship between markups and the number of customers is negative, but an order of magnitude smaller in size. Column (2) includes UPC-, market-, and retailer-fixed effects interacted with year fixed effects, and the importance of the sales per customer for markup measure remains (while the coefficient on the number of customers becomes even smaller). Column (3) includes a set of UPC-market-year fixed effects that absorb any differences in marginal costs at those levels (including the variation in wholesale costs), and columns (4), (5), and (6) allow for additional sets of fixed effects. Once we exploit our geographical variation and include retailer-year-UPC fixed effects, the coefficient on the number of customers declines to -0.013, while the coefficient on the average sales per customer remains large at 0.171. Finally, our most conservative specification in Column (6) shows that there is still a positive and significant relationship between markups and sales per customer, but no significant relationship with the number of customers (the coefficient is precisely estimated close to zero). Table B.3 replaces the number of customers in Table B.2 with total sales, and it still shows that the relevant margin for the relationship between size and markups is the average sales per customer.

Table B.2: UPC-market-retailer-year-level analysis

	ln $\text{Markup}_{\text{urmt}}$					
	(1)	(2)	(3)	(4)	(5)	(6)
ln $p_{\text{urmt}}q_{\text{urmt}}$	0.550***	0.186***	0.270***	0.171***	0.169***	0.187***
	(0.035)	(0.027)	(0.041)	(0.021)	(0.019)	(0.019)
ln m_{urmt}	-0.063***	-0.040***	-0.063***	-0.013***	-0.011***	-0.004
	(0.018)	(0.005)	(0.021)	(0.004)	(0.003)	(0.002)
Observations	426032	426032	426032	426032	426032	426032
R^2	0.126	0.550	0.599	0.844	0.854	0.928
UPC-year FE		✓				
market-year FE		✓				
retail-year FE		✓				
UPC-market-year FE			✓	✓	✓	✓
retail-year-UPC FE				✓	✓	✓
year-market-retail FE					✓	✓
UPC-market-retail FE						✓

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$; standard errors are three-way clustered by product group, retail, and market. The Markup_{urmt} is measured as the difference between retail-level price and wholesale cost, $p_{urmt}q_{urmt}$ is the sales per customer, and m_{urmt} is the number of customers, where the subscript u is UPC, r is retail, m is market, and t is year. All variables are projection-factor adjusted. We balance the sample across columns based on the tightest specification in column (6); our final sample includes 7,802 number of UPCs, 45 markets, 6 years, and 167 retailers.

Table B.3: UPC-market-retailer-year-level analysis

	ln Markup$_{urmt}$					
	(1)	(2)	(3)	(4)	(5)	(6)
ln $p_{urmt}q_{urmt}$	0.612***	0.226***	0.334***	0.184***	0.181***	0.191***
	(0.046)	(0.028)	(0.046)	(0.022)	(0.020)	(0.020)
ln S_{urmt}	-0.063***	-0.040***	-0.063***	-0.013***	-0.011***	-0.004
	(0.018)	(0.005)	(0.021)	(0.004)	(0.003)	(0.002)
Observations	426032	426032	426032	426032	426032	426032
R^2	0.126	0.550	0.599	0.844	0.854	0.928
UPC-year FE		✓				
market-year FE		✓				
retail-year FE		✓				
UPC-market-year FE			✓	✓	✓	✓
retail-year-UPC FE				✓	✓	✓
year-market-retail FE					✓	✓
UPC-market-retail FE						✓

Notes: The regression specification is the same as what is used in Table B.2 except that we replace the number of customers with total sales.

B.1.4 Sales and SGA Expenses: Decomposition by Durability of Products

By using the same durability measure we used in Figure B.3, we analyze the possibility of different correlation of SGA and Sales based on the product durability in Table B.4. The correlation is not different across the durability of products.

Table B.4: Sales and SGA Expenses: Decomposition by Durability of Products

	Decomposition of ln S_{igt}			ln m_{igt}: New vs. Old	
	(1) ln S	(2) ln pq	(3) ln m	(4) ln m^{New}	(5) ln m^{Old}
ln SGA$_{it}$	0.079**	-0.006	0.085***	0.078***	0.024
	(0.036)	(0.017)	(0.030)	(0.030)	(0.038)
ln SGA$_{it}$ x Durability	0.040	0.026	0.013	0.042	-0.018
	(0.046)	(0.028)	(0.032)	(0.039)	(0.046)
Observations	13131	13131	13131	13131	13131
R^2	0.962	0.909	0.965	0.943	0.961
Firm-year Controls	✓	✓	✓	✓	✓
Group-year FE	✓	✓	✓	✓	✓
SIC-year FE	✓	✓	✓	✓	✓
Group-firm FE	✓	✓	✓	✓	✓

Notes: The regression specification is the same as what is used in Table 4 except that we additionally include the durability measure and its interaction with the log SGA.

B.1.5 Semi-variable Nature of SGA

This section establishes that the SGA expenses have a semi-variable nature and are correlated with firms' sales in the short-run. Previous studies that examined the non-production cost of firms (SGA) made a polar opposite assumption on the variable nature of this cost. For instance, in measuring price-cost markups, Traina (2019) includes non-production costs as variable costs, whereas De Loecker et al. (2020) interprets non-production costs as fixed costs in their baseline approach. We empirically assess such assumption by comparing the co-movement of sales and SGA with the co-movement of sales and other costs that are commonly assumed to be variable and fixed in the short-run in the literature: COGS and capital.

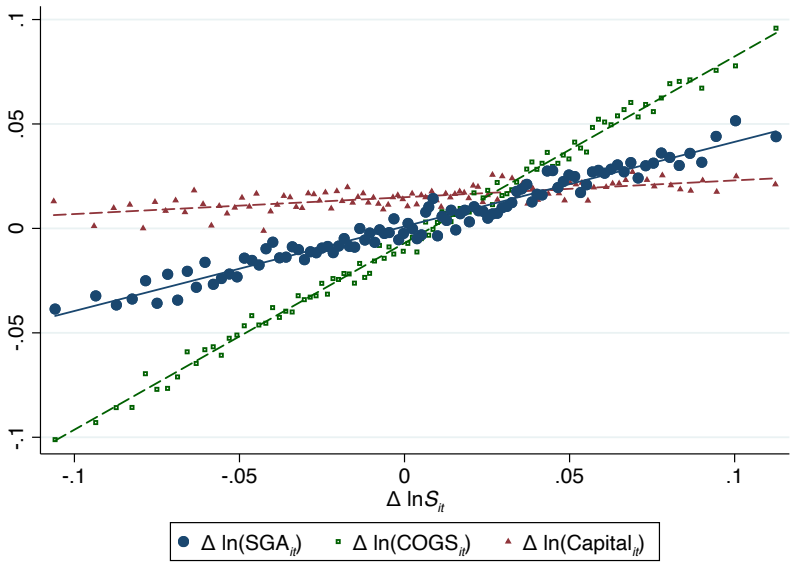


Figure B.4: The Semi-variable Nature of SGA

Notes: The figure shows the binned scatter plot of the correlation between quarterly change in log sales and quarterly change in: i) log SGA, ii) log COGS, and iii) log capital for firms in the quarterly Compustat dataset. We also plot the best linear fit for each variable. The correlations control for year and firm fixed effects. We restrict the sample to observations with a change in the log of sales was between -0.1 and +0.1 (the 25th and 75th percentiles of the quarterly change of log sales is -8% and 11%, respectively). We adapt perpetual inventory method following Traina (2019) and use Gross and NET Capital (PPEGT and PPENT) and deflate investment with NIPA's non-residential fixed investment good deflator to measure the capital. There are 17,168 firms in 1964-2016 used in this analysis.

Our results suggest that the SGA expenses have both variable and fixed components; it is more variable than the capital expenditure but is less variable than the COGS. Figure B.4 reports the binned scatter plot of changes in sales against changes if firm's costs for a range of $\Delta \ln S_{i,t}$ between -10% and 10%, which are approximately the 25th and 75th percentiles of the $\Delta \ln S_{it}$ distribution. Consistent with the view in the literature (e.g., Edmond et al. 2018; De Loecker et al. 2020), sales exhibit the largest co-movement with production costs ($\beta = 0.894$; SE 0.008), and the lowest co-movement with investment ($\beta = 0.081$; SE 0.008).

We consider other empirical specifications to confirm the semi-variable nature of SGA expenses. Table B.5 presents the regression results that correspond to Figure B.4. The semi-variable nature of SGA is clear in this Table with and without fixed effects. We also show

that R&D is more variable than capital but is not as variable as SGA. Figure B.5 replicates Figure B.4 by using the full sample, and the semi-variability of SGA expense remains strong. Figure B.6 presents the cross-correlation, which further supports the short-run variability of SGA. Although there is a strong contemporaneous correlation of SGA and sales, SGA is generally not correlated with the forward or backward sales.

Table B.5: The Semi-variable Nature of SGA

	(1) $\Delta ln(\text{COGS}_{it})$	(2) $\Delta ln(\text{COGS}_{it})$	(3) $\Delta ln(\text{SGA}_{it})$	(4) $\Delta ln(\text{SGA}_{it})$	(5) $\Delta ln(\text{Capital}_{it})$	(6) $\Delta ln(\text{Capital}_{it})$	(7) $\Delta ln(\text{R\&D}_{it})$	(8) $\Delta ln(\text{R\&D}_{it})$
ΔlnS_{it}	0.920***	0.894***	0.473***	0.405***	0.133***	0.081***	0.278***	0.200***
	(0.008)	(0.008)	(0.009)	(0.010)	(0.008)	(0.008)	(0.028)	(0.031)
R^2	0.055	0.162	0.011	0.103	0.002	0.155	0.002	0.083
Year FE	No	Yes	No	Yes	No	Yes	No	Yes
Firm FE	No	Yes	No	Yes	No	Yes	No	Yes
N	293962	292739	292785	291547	134743	133745	69845	69363

Notes: The dependent variables are the quarterly change in log COGS, SGA and Capital. The estimation method used in all columns is OLS. Standard errors (in parentheses) are clustered at the firm level.

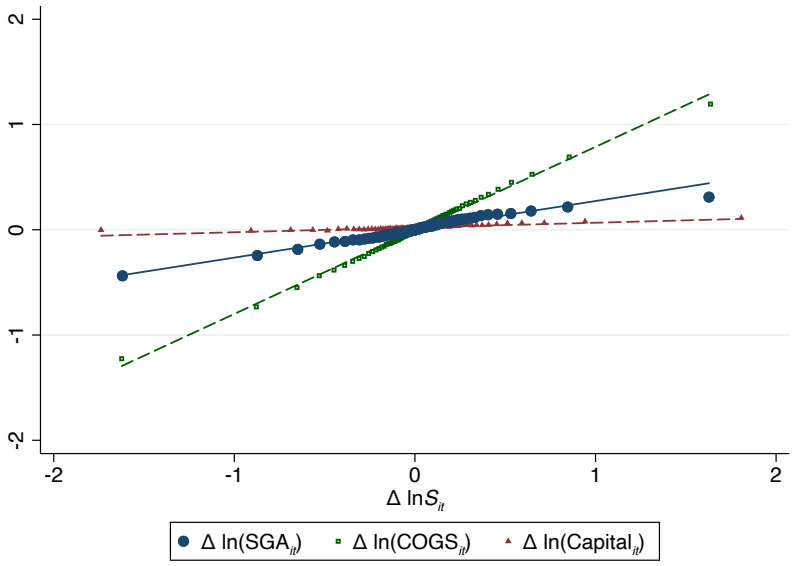
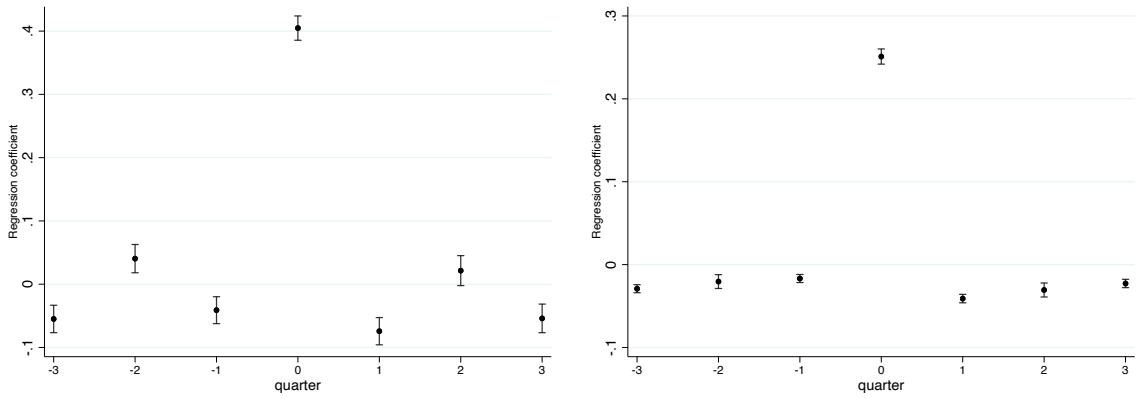


Figure B.5: The Semi-variable Nature of SGA

Notes: The figure shows the binned scatter plot of the correlation between quarterly change in log sales and quarterly change in: i) log capital, ii) log COGS, and iii) log SGA, for firms in the quarterly Compustat dataset. We also plot the best linear fit for each variable. The correlations control for firm and quarter fixed effects.

Figure B.6: Cross-correlation



(a) Trimmed Sample (b) Full Sample

Notes: Figure B.6a uses the trimmed sample presented in the main body of the paper, and Figure B.6b uses the full sample. 95% confidence intervals are presented for every estimate. Figure B.6a replicates the column (4) in Table B.5 at quarter = 0.

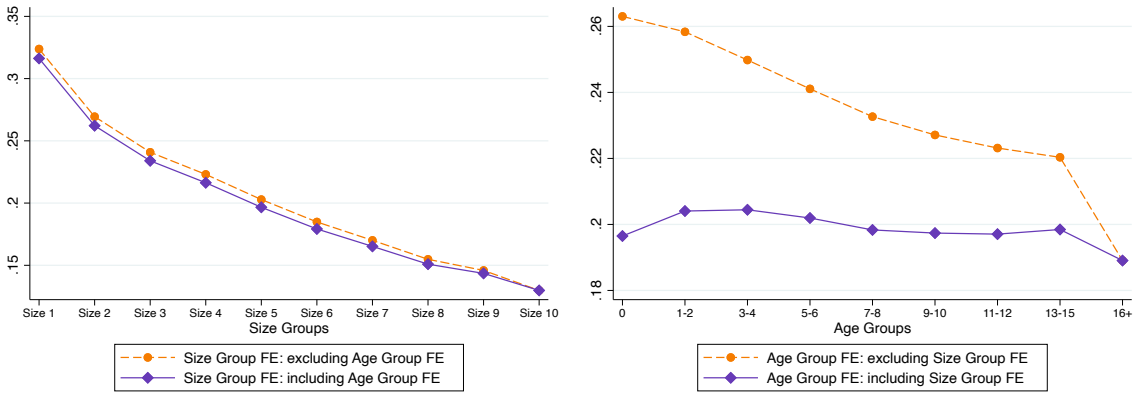
B.2 SGA and firm heterogeneity

This section analyzes the potential characteristics associated with firms' non-production costs to understand the underlying firm heterogeneity. We focus on two characteristics of firms, size and age, since these two are the most notably known to influence firm-level outcomes in macroeconomics literature (see, e.g., Hopenhayn, Neira and Singhania, 2018; Autor et al., 2020). We use the following regression specification to study the correlation of non-production costs with firm size and age:

$$y_{it} = \sum_{m=1}^{10} \alpha_m \mathbf{1}(\text{size group}_{it} = m) + \sum_{a=1}^{9} \alpha_a \mathbf{1}(\text{age group}_{it} = a) + \alpha_s + \alpha_t + \varepsilon_{it}$$

where i is firm and t is year. y_{it} is the SGA-to-OPEX ratio, α_s is 1-digit SIC sector fixed effect, and α_t is year fixed effect. Our primary interest is α_m and α_a, which measure the average SGA-to-OPEX ratio for each size or age group conditional on the other firm characteristic. There are 10 size groups based on the decile of the distribution of firms' market share in each year and 9 age groups $(0, 1\text{-}2, 3\text{-}4, 5\text{-}6, 7\text{-}8, 9\text{-}10, 11\text{-}12, 13\text{-}15, 16+)$ following Haltiwanger et al. (2013).[29] The sector and year fixed effects are included to analyze the average SGA-to-OPEX within each sector and each year.

Figure B.7: SGA Share of OPEX: Size vs. Age



(a) Size Profile

(b) Age Profile

Notes: Figure B.7a plots the size profiles of the COGS-to-OPEX ratio (α_m) and Figure B.7b plots the age profiles of the SGA-to-OPEX ratio (α_a) by estimating equation B.2. The baseline group is the largest group or the oldest group; we plot the average SGA-to-OPEX for the baseline group and add the estimated coefficient of indicator variable for each of the other groups. The size groups are based on the decile of the distribution of firms' market share in each year and the age groups are $(0, 1\text{-}2, 3\text{-}4, 5\text{-}6, 7\text{-}8, 9\text{-}10, 11\text{-}12, 13\text{-}15, 16+)$ following Haltiwanger et al. (2013). In Figure B.7a, the dotted orange line plots the size group fixed effects without the inclusion of the age group fixed effects in the regression, and the solid purple line plots the size group fixed effects with the inclusion of the age group fixed effects. Similarly, in Figure B.7b, the dotted orange line plots the age group fixed effects without the inclusion of the size group fixed effects in the regression, and the solid purple line plots the age group fixed effects with the inclusion of the size group fixed effects.

[29] Following Haltiwanger et al. (2013), we exclude the first fifteen years of the dataset for all analyses using age, and we group firms older than sixteen years, because we do not know for certain a firm's IPO date for firms that were in the Compustat data since the beginning of the sample.

We find that small firms have substantially larger non-production expenses compared to large firms, as presented in Figure B.7. Figure B.7a shows that smallest firms spend more than 30% of their expenses on non-production, while the largest firms spend approximately 20 percent points less than the smallest firms. Although there is a similarly large heterogeneity of non-production expenses across firm ages, it is largely driven by the firm size heterogeneity; conditioning on firm size group fixed effects, Figure B.7b shows that there is a negligible difference in the share of non-production expenses across old and young firms.

C Model with Heterogeneous Tastes and Sorting

In this section, we provide two model extensions to our households' preferences. In the first, we allow for idiosyncratic preference shocks that affect the optimal quantities consumed by each consumer. In the second, we allow for shifters that affect the perceived "quality" of each variety. We discuss how each extension affects the total demand of a firm as well as the relationship between number of customers and average sales per customer. We also provide empirical evidence for why we abstract away from these extensions in the main analysis.

C.1 Taste for Quantity

Consider the following extension of the Kimball aggregator in Equation (3.1):

$$\int_0^{N_t} \int_0^1 \mathbf{1}_{\{j \in m_{i,t}\}} \xi_{i,j,t}^{-1} \Upsilon \left(\frac{\xi_{i,j,t} c_{i,j,t}}{C_t} \right) djdi = 1,$$

where $\xi_{i,j,t}$ is now household j's "quantity" taste for variety i at time t. The implied relative demand for firm i at t is then given by

$$\frac{c_{i,t}}{C_t} = \underbrace{m_{i,t}}_{\text{number of customers}} \times \underbrace{\mathbb{E}[\xi_{i,j,t} | j \in m_{i,t}] \Upsilon'^{-1} \left(\frac{p_{i,t}}{D_t} \right)}_{q_{i,t} \equiv \text{ demand per customer}},$$

where $\mathbb{E}[\xi_{i,j,t} | j \in m_{i,t}]$ is the average quantity taste of firm i's customers. The introduction of such shocks will affect firm i's demand depending on the nature of sorting between customers and firms. If there is no sorting or selection in who is matched to a firm, then $\mathbb{E}[\xi_{i,j,t} | j \in m_{i,t}] = \mathbb{E}[\xi_{i,j,t}]$ and this extension replicates the demand function in the main text. However, sorting generates a correlation between the number of customers and average sales per customer. For example, with positive sorting, firms with a larger customer base should sell less per customer on average (because the marginal customer always buys less than the average customer).[30] Instead, with negative sorting (e.g., when marginal consumers are the ones who experiment with the product and buy more than the average con-

[30] For illustration purposes, we assume for any i and t that $\xi_{i,j,t}$ are i.i.d. draws from a Pareto distribution with shape parameter $\theta > 0$ and normalized so that its unconditional mean is 1 ($\mathbb{E}[\xi_{i,j,t}] = 1$). Let us consider the extreme case of perfect sorting, where the customers that are matched to a firm are the ones who have the highest taste for its product. With this natural assumption, for a given mass of customers $m_{i,t}$, the firm's customers are the household members for whom $\xi_{i,j,t} \geq (1 - \theta^{-1}) m_{i,t}^{-\theta^{-1}}$ and the selection term is $\mathbb{E}[\xi_{i,j,t} | j \in$

sumer), firms with a larger customer base should sell more per customer on average. More generally, this model implies the following relationship between average sales per customer and a firm's number of customers

$$\ln(p_{i,t}q_{i,t}) = \beta_0 \ln(m_{i,t}) + \ln\left(p_{i,t}\Upsilon'^{-1}\left(\frac{p_{i,t}}{D_t}\right)\right),$$

where the sign and magnitude of β_0 determine the type and strength of sorting, respectively. To test this, we aggregate the Nielsen Homescan Panel data at the UPC-year-level. For this analysis, we use the whole sample available in Homescan Panel data. We compute the price of each product as sales divided by the unit-adjusted quantity. Table C.1 shows the results of OLS regressions of the log average sales per customer of UPC u at time t on the log number of customers and polynomials of log price that approximate the nonlinear function $\Upsilon'^{-1}(\cdot)$. Across all specifications, we consistently find a small and positive relationship between the size of the customer base and average sales per customer. Firms with a 1% higher number of customers sell on average 0.03% more per customer. Given the economic insignificance of this estimate, we do not consider this kind of preference heterogeneity in our model, which as discussed below is a conservative choice.

Table C.1: Average Sales per Customer and The size of the Customer Base

	$\ln p_{ut}q_{ut}$				
	(1)	(2)	(3)	(4)	(5)
$\ln m_{ut}$	0.026***	0.025***	0.023**	0.029***	0.029***
	(0.009)	(0.009)	(0.009)	(0.002)	(0.002)
$\ln p_{ut}$	0.162***	0.180***	0.132***	0.680***	0.688***
	(0.028)	(0.037)	(0.027)	(0.017)	(0.017)
$\ln p^2_{ut}$		0.021*	0.025	-0.003	-0.003
		(0.012)	(0.015)	(0.005)	(0.005)
$\ln p^3_{ut}$			0.005	-0.001	-0.001
			(0.003)	(0.001)	(0.001)
Observations	9097076	9097076	9097076	8452707	8452707
R^2	0.086	0.096	0.102	0.851	0.852
UPC FE				✓	✓
Year FE				✓	
Product Group-Year FE					✓

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$; Standard errors are clustered by product group. The variable $p_{ut}q_{ut}$ denotes the sales per customer of UPC u at time t, m_{ut} is the number of customers, and p_{ut} is the price.

Implications of Sorting for Misallocation Results A potential concern in our analysis of efficiency losses from the misallocation of demand is that sorting of customers might exacerbate or reduce welfare losses from misallocation. For example, since with positive sorting the marginal customer values a variety less than the average customer, the marginal value of allocating customers to more productive firm might not be as high as in our baseline model. However, regression results in Table C.1 show that, if anything, there is negative, albeit small, sorting (i.e., the marginal customer buys more than the average customer).

$m_{i,t}] = m_{i,t}^{-\theta^{-1}}$. In this case, $\beta_0 = -\theta^{-1}$.

Viewed through the lens of this model extension, these results indicate that the quantified welfare losses from misallocation of demand in Section 6 provide a *lower* bound on the actual welfare losses once this small negative sorting is taken into account.

C.2 Taste for Quality

Another source of preference heterogeneity can be the different perception of the "quality" of a variety across customers. For this, consider the following extension of the Kimball aggregator in Equation (3.1):

$$\int_0^{N_t} \int_0^1 \mathbf{1}_{\{j \in m_{i,t}\}} \zeta_{i,j,t} \Upsilon \left(\frac{c_{i,j,t}}{C_t} \right) djdi, = 1$$

where $\zeta_{i,j,t}$ is now the household j's "quality" taste for variety i at time t. We assume for any i and t that $\zeta_{i,j,t}$ is an i.i.d. and it distribution is scaled so that its unconditional mean is 1 ($\mathbb{E}[\zeta_{i,j,t}] = 1$). The implied relative demand of firm i is then given by

$$q_{i,t} \equiv \frac{c_{i,t}}{C_t} = \int_0^1 \mathbf{1}_{\{j \in m_{i,t}\}} \Upsilon'^{-1} \left(\frac{p_{i,t}}{\zeta_{i,j,t} D_t} \right) dj$$

The curvature of Υ'^{-1}, and thus the elasticity of demand, now interacts with the distribution of $\zeta_{i,j,t}$, but we can show that up to first order approximation, sorting would imply a correlation between average demand per customer and the number of customers, similar to the quantity shocks. To see this, note that up to first order approximation around the point $\frac{p_{i,t}}{\mathbb{E}[\zeta_{i,j,t}] D_t}$:

$$\Upsilon'^{-1} \left(\frac{p_{i,t}}{\zeta_{i,j,t} D_t} \right) \approx \Upsilon'^{-1} \left(\frac{p_{i,t}}{D_t} \right) \left(1 + \sigma_{i,t} (\zeta_{i,j,t} - 1) \right)$$

where $\sigma_{i,t}$ is the elasticity of demand evaluated at average quality taste, $p_{i,t}/D_t$, and hence depends only on price. Now, using this approximation we can write demand as

$$\frac{c_{i,t}}{C_t} \approx \underbrace{m_{i,t}}_{\text{number of customers}} \times \underbrace{\Upsilon'^{-1} \left(\frac{p_{i,t}}{D_t} \right) \left(1 + \sigma_{i,t} (\mathbb{E}[\zeta_{i,j,t} | j \in m_{i,t}] - 1) \right)}_{q_{i,t} \equiv \text{demand per customer}}$$

where $\mathbb{E}[\zeta_{i,j,t} | j \in m_{i,t}]$ is the average quality taste of firm i's customers. Similar to quantity shocks, if there is positive (negative) sorting, then $\mathbb{E}[\zeta_{i,j,t} | j \in m_{i,t}]$ is a decreasing (increasing) function of $m_{i,t}$ and we should see a negative (positive) relationship between $q_{i,t}$ and $m_{i,t}$, which brings us to the same argument in the previous section for quantity shocks.

D Derivations and Proofs

D.1 Lemma 1

Proof. We start from the expression for the optimal price of the firm:

$$\ln\left(\frac{p_{i,t}}{D_t}\right) = \ln(\varepsilon_{i,t}) - \ln(\varepsilon_{i,t}-1) + \ln\left(\frac{mc_{i,t}}{D_t}\right), \tag{D.1}$$

where

$$\varepsilon_{i,t} = -\frac{\partial \ln(q_{i,t})}{\partial \ln(p_{i,t})} = \frac{\sigma}{1-\eta\ln(p_{i,t})+\eta\ln(D_t(1-\sigma^{-1}))}. \tag{D.2}$$

The last equality in Equation (D.2) follows from the expression of demand per match in Equation (3.4). Differentiating Equation (D.1) we have:

$$d\ln\left(\frac{p_{i,t}}{D_t}\right) = (1-\mu_{i,t})d\ln(\varepsilon_{i,t}) + d\ln\left(\frac{mc_{i,t}}{D_t}\right) = \frac{1}{1+\eta\sigma^{-1}\varepsilon_{i,t}(\mu_{i,t}-1)}d\ln\left(\frac{mc_{i,t}}{D_t}\right),$$

where $\mu_{i,t} \equiv \frac{\varepsilon_{i,t}}{\varepsilon_{i,t}-1}$ is the firm's markup. Then, it follows that

$$d\ln(\mu_{i,t}) = d\ln(p_{i,t}) - d\ln(mc_{i,t}) = -\frac{\eta\sigma^{-1}\varepsilon_{i,t}(\mu_{i,t}-1)}{1+\eta\sigma^{-1}\varepsilon_{i,t}(\mu_{i,t}-1)}d\ln\left(\frac{mc_{i,t}}{D_t}\right).$$

□

D.2 Proposition 1

Proof. Consider the sales per match of firm i normalized by the demand index D_t, $p_{i,t}q_{i,t}/D_t$. Differentiating the log of this quantity, we have

$$d\ln\left(\frac{p_{i,t}q_{i,t}}{D_t}\right) = (1-\varepsilon_{i,t})d\ln\left(\frac{p_{i,t}}{D_t}\right) = -\frac{\varepsilon_{i,t}-1}{1+\eta\sigma^{-1}\varepsilon_{i,t}(\mu_{i,t}-1)}d\ln\left(\frac{mc_{i,t}}{D_t}\right).$$

Therefore, combining this expression with Equation (3.14), we have

$$d\ln(\mu_{i,t}) = \eta\sigma^{-1}\mu_{i,t}(\mu_{i,t}-1)d\ln\left(\frac{p_{i,t}q_{i,t}}{D_t}\right).$$

□

D.3 Corollary 1

Proof. We show this result by differentiating the markup of the firm, while keeping the total level of firms' sales constant. The latter implies that

$$0 = d\ln(p_{i,t}y_{i,t}) = d\ln(p_{i,t}q_{i,t}) + d\ln(m_{i,t}) \Rightarrow d\ln(p_{i,t}q_{i,t}) = -d\ln(m_{i,t}).$$

Now, plugging this into Equation (3.15) we have

$$d\ln(\mu_{i,t}) = -\eta\sigma^{-1}\mu_{i,t}(\mu_{i,t}-1)d\ln\left(\frac{m_{i,t}}{D_t}\right).$$

Notice that in deriving this expression we have only used the elasticity of markups with respect to the marginal cost of the firm, keeping its productivity within a period fixed. Therefore, this result only depends on the fact that, keeping sales and productivity fixed, firms with higher customer-base have higher marginal costs and hence by Lemma 1 lower markups. □

D.4 Proposition 2

Proof. This relationship is obtained directly from the first order condition of the firm's problem with respect to $m_{i,t}$. For the rest of the proof, we derive this first order condition.

We start by showing that the firm's customer acquisition constraint always binds (meaning that the firm never disposes their existing customers). To show this, note that it cannot be the case that $l_{i,s,t} > 0$ but $m_{i,t} < (1-\delta)m_{i,t-1} + \frac{l^{\phi}_{i,s,t}}{P_{m,t}}$ since the firm can keep the same $m_{i,t}$ with a lower $l_{i,s,t}$. Thus, if $m_{i,t} < (1-\delta)m_{i,t-1} + \frac{l^{\phi}_{i,s,t}}{P_{m,t}}$, then optimality requires that $l_{i,s,t} = 0$. Now, suppose $l_{i,s,t} = 0$ but $m_{i,t} < (1-\delta)m_{i,t-1}$. Note that in this case, the slope of the firm's "production" profit function with respect to $m_{i,t}$ is given by

$$\frac{\partial}{\partial m_{i,t}}(p_{i,t}y_{i,t} - W_t l_{i,p,t}) = (p_{i,t} - mc_{i,t})\frac{y_{i,t}}{m_{i,t}} > 0,$$

where the last equality follows from the fact that for any choice of $q_{i,t} > 0$, the firm's markup is always strictly larger than 1 and hence $p_{i,t} > mc_{i,t}$. Thus, the firm's profit is strictly increasing in $m_{i,t}$ and since $m_{i,t} < (1-\delta)m_{i,t-1}$, then the firm can increase its $m_{i,t}$ at no cost and gain more profits at time t. Moreover, this will not affect firms' profits in the future since the firms can always dispose of the increase in $m_{i,t}$ in the next period at no cost. Hence, optimality requires that $m_{i,t} = (1-\delta)m_{i,t-1} + \frac{l^{\phi}_{i,s,t}}{P_{m,t}}$.

Now, in writing firm i's problem at time t, replace $l_{i,p,\tau} = (y_{i,\tau}/z_{i,\tau})^{\alpha^{-1}}$, $y_{i,\tau} = m_{i,\tau}q_{i,\tau}C_\tau$, and $l_{i,s,\tau} = P^{\phi^{-1}}_{m,\tau}(m_{i,\tau} - (1-\delta)m_{i,\tau-1})^{\phi^{-1}}$ to obtain the problem as

$$\max_{\{p_{i,\tau},m_{i,\tau},q_{i,\tau}\}_{\tau\geq t}} \mathbb{E}_t \sum_{\tau\geq t}(\beta\nu)^{\tau-t}C_\tau^{-\gamma}\left(\prod_{h=t}^{\tau}\mathbf{1}_{i,h}\right)\times$$
$$\left[p_{i,\tau}m_{i,\tau}q_{i,\tau}C_\tau - W_\tau\left(\frac{m_{i,\tau}q_{i,\tau}C_\tau}{z_{i,\tau}}\right)^{\alpha^{-1}} - W_\tau P^{\phi^{-1}}_{m,\tau}(m_{i,\tau}-(1-\delta)m_{i,\tau-1})^{\phi^{-1}} - W_\tau\chi\right]$$
$$s.t. \quad q_{i,\tau} = \left[1-\eta\ln\left(\frac{p_{i,\tau}}{D_\tau(1-\sigma^{-1})}\right)\right]^{\frac{\sigma}{\eta}}.$$

Next, if $\mathbf{1}_{i,t} = 0$, then $l_{i,s,t} = 0$. However, conditional on $\mathbf{1}_{i,t} = 1$, the FOC with respect to $m_{i,t}$ is

$$\begin{aligned} 0 =& \mathbf{1}_{i,t}(p_{i,t} - \alpha^{-1}\frac{W_t l_{i,p,t}}{y_{i,t}})q_{i,t}C_t - \mathbf{1}_{i,t}\phi^{-1}\frac{W_t l_{i,s,t}}{m_{i,t} - (1-\delta)m_{i,t-1}} \\ &+ \beta\nu(1-\delta)\mathbb{E}_t\left[\left(\frac{C_{t+1}}{C_t}\right)^{-\gamma}\mathbf{1}_{i,t+1}\phi^{-1}\frac{W_{t+1}l_{i,s,t+1}}{m_{i,t+1} - (1-\delta)m_{i,t}}\right]. \end{aligned}$$

Replacing $mc_{i,t} = \alpha^{-1}\frac{W_t l_{i,p,t}}{y_{i,t}}$ and iterating the FOC forward gives us the expression of interest. □

D.5 Planner's Problem

Proof. The Planner's problem for this economy is given by

$$\max_{\left\{\substack{(c_{i,j,t})_{j\in[0,1]},(\mathbf{1}_{i,t})_{i\in N_{t-1}\cup\Lambda_t}, \\ (\delta_{i,t},m_{i,t},l_{i,p,t},l_{i,s,t})_{i\in N_t},C_t}\right\}_{t\geq 0}} \sum_{t=0}^{\infty}\beta^t\left[\frac{C_t^{1-\gamma}}{1-\gamma} - \xi\frac{L_t^{1+\psi}}{1+\psi}\right] \tag{D.3}$$

subject to

$$\int_0^1 \mathbf{1}_{\{j\in m_{i,t}\}}c_{i,j,t}dj = z_{i,t}l_{i,p,t}^{\alpha}, \quad \forall i \in N_t, \tag{D.4}$$

$$\int_{i\in N_t}\int_0^1 \mathbf{1}_{\{j\in m_{i,t}\}}\Upsilon\left(\frac{c_{i,j,t}}{C_t}\right)djdi = 1 \tag{D.5}$$

$$\int_{i\in N_t}(l_{i,p,t} + l_{i,s,t} + \chi)di = L_t \tag{D.6}$$

$$\int_{i\in N_t} m_{i,t}di = 1 \tag{D.7}$$

$$N_t = \{i \in N_{t-1}\cup\Lambda_t : \mathbf{1}_{i,t}\nu_{i,t} = 1\}, \quad N_{-1} \text{ given.} \tag{D.8}$$

$$m_{i,t} = (1-\delta_{i,t})m_{i,t-1} + \frac{1 - \int_{i\in N_t}(1-\delta_{i,t})m_{i,t-1}}{\int_{i\in N_t} l_{i,s,t}^{\phi}di}l_{i,s,t}^{\phi}, \quad \forall i \in N_t, \tag{D.9}$$

$$\delta_{i,t} \in [\delta, 1], l_{i,s,t} \geq 0, \quad \forall i \in N_t, \tag{D.10}$$

$$\int_{i\in N_t} l_{i,s,t}di = \bar{L}_{s,t} > 0. \tag{D.11}$$

Here, Equation (D.4) requires that for every firm, their supply meets their allocated demand, Equation (D.5) is the Kimball aggregator that implicitly defines C_t given the planner allocation of demand, Equation (D.6) requires that labor supply meets demand for labor from production, advertisement and overhead costs, Equation (D.7) requires that the matching market clears, Equation (D.8) is the law of motion for the set of operating firms given an entry/exit policy by the planner, Equation (D.9) determines firm i's evolution of customers

given an allocation for advertising, Equation (D.10) requires the non-negativity of labor for advertising and the constraint that while the planner can separate customers from firms the separation rate should be at least δ, and finally Equation (D.11) requires that the planner at least spends $\bar{L}_{s,t}$ on advertisement. This last constraint is for an arbitrary but a strictly positive $\bar{L}_{s,t}$—in Lemma 2 we show that the level of this quantity does not matter for the optimal distribution of customers, which is also well-defined in the limit when $\bar{L}_{s,t} \to 0$. □

D.6 Lemma 2

Proof. Suppose that at any given time t, a choice for N_t is fixed. Suppose next that the planner desires to allocate matches according to a rule

$$\mathcal{A} : (i \to m^*_{i,t})_{i \in N_t}.$$

Note that this can be any arbitrary allocation of matches as long as it is feasible:

$$m^*_{i,t} \geq 0, \quad \forall i \in N_t$$
$$\int_{i \in N_t} m^*_{i,t} di = 1.$$

To show that the allocation $\mathcal{A}$ is implementable on N_t for any given level of $\bar{L}_{s,t}$, we need to show that (1) it is generated by a choice of $(\delta \leq \delta_{i,t} \leq 1, l_{i,s,t} \geq 0)_{i \in N_t}$, and (2) it is feasible $\int_{i \in N_t} l_{i,s,t} di = \bar{L}_{s,t}$. We show this by construction. In particular, consider the choice:

$$\left(\delta^*_{i,t} = 1, l^*_{i,s,t} = \bar{L}_{s,t} \frac{m^{*\phi^{-1}}_{i,t}}{\int_{i \in N_t} m^{*\phi^{-1}}_{i,t} di} \right)_{i \in N_t}$$

That is, first, let the planner separate all the matches from their corresponding firms ($\delta^*_{i,t} = 1$) and then reallocate them based on $\mathcal{A}$. It follows that conditions (1) and (2) from above hold by construction. Next, to verify that these values implement $\mathcal{A}$, observe that

$$m_{i,t} \equiv (1 - \delta_{i,t}) m_{i,t-1} + \left(1 - \int_{i \in N_t} (1 - \delta_{i,t}) m_{i,t-1} di \right) \frac{l^{*\phi}_{i,s,t}}{\int_{i \in N_t} l^{*\phi}_{i,s,t} di} = m^*_{i,t}.$$

Finally, to confirm feasibility, note that $\int_{i \in N_t} l^*_{i,s,t} = \bar{L}_{s,t}$. □

D.7 Proposition 3

Proof. The results in this Proposition follow from the first order conditions of the planner's problem in Equation (D.3), fixing the planner's other choices at an arbitrary allocation. In the remainder of this proof, we characterize these first order conditions.

Formally, for firm i and period t, let $\beta^t \varphi_{c,i,t} di$ be the shadow cost on Equation (D.4); for period t, let $\beta^t \varphi_{Y,t}$, $\beta^t \varphi_{L,t}$ and $\beta^t \varphi_{m,t}$ be the shadow costs on Equation (D.5), Equation (D.6) and Equation (D.7), respectively. Moreover, similar to the equilibrium allocation, let us

define $q_{i,j,t} \equiv \frac{c_{i,j,t}}{C_t}$. It is straight forward to show that for $j \notin m_{i,t}$, $q_{i,j,t} = 0$. So from here on forward we only refer to $q_{i,j,t}$ when $j \in m_{i,t}$.

The first order conditions with respect to $q_{i,j,t}$ are

$$\varphi_{c,i,t} C_t = \Upsilon'(q_{i,j,t}) \varphi_{\Upsilon,t}, \quad \forall j \in m_{i,t}. \tag{D.12}$$

It immediately follows that all households that are matched to a variety consume the same amount:

$$q_{i,j,t} = q_{i,t}, \quad \forall j \in m_{i,t}. \tag{D.13}$$

Replacing this result in Equation (D.4) and Equation (D.5) and taking the first order condition with respect to $m_{i,t}$, we have:

$$\varphi_{c,i,t} q_{i,t} C_t + \varphi_{m,t} = \Upsilon(q_{i,t}) \varphi_{\Upsilon,t}. \tag{D.14}$$

Notice that in deriving this first order condition, we have ignored the constraint in Equation (D.9). The reason that we can do this goes back to Lemma 2, which states that any choice of $(m_{i,t})_{i \in N_t}$ can be implemented without any loss of generality. Therefore, we can ignore the constraint in Equation (D.9) and then use Lemma 2 to show that it is satisfied.

Next, replacing Equation (D.13) in Equation (D.12), multiplying it by $q_{i,t}$ and subtracting it from Equation (D.14), we have:

$$\varphi_{m,t} = \left[\Upsilon(q_{i,t}) - q_{i,t} \Upsilon'(q_{i,t})\right] \varphi_{\Upsilon,t}.$$

Since $\varphi_{\Upsilon,t} \neq 0$,[31] it follows that

$$\Upsilon(q_{i,t}) - q_{i,t} \Upsilon'(q_{i,t}) = \frac{\varphi_{m,t}}{\varphi_{\Upsilon,t}}, \quad \forall i \in N_t.$$

Notice that the left hand side of this equation is only a function of $q_{i,t}$ and it is strictly monotonic in $q_{i,t} > 0$.[32] Moreover, the right hand side of the equation is only a function of time-t shadow costs and is independent of i. Hence, there exists a unique q_t^* such that

$$q_{i,t} = q_t^* \quad \forall i \in N_t.$$

Replacing this last equation into Equation (D.5) we have:

$$\int_{i \in N_t} m_{i,t} \Upsilon(q_t^*) di = 1 \Rightarrow \Upsilon(q_t^*) = 1 \Rightarrow q_t^* = 1,$$

where the second statement uses the market clearing condition for matches in Equation (D.7) and the last statement uses the strict monotonicity of $\Upsilon(x)$ and the fact that $\Upsilon(1) = 1$.

[31] To see why, suppose not. Then, by Equation (D.12), either $C_t = 0$—which is clearly not optimal since marginal utility approaches infinity as $C_t \to 0$—or $\varphi_{c,i,t} = 0$—which means that the household can freely supply infinite labor to firm i at t and is also a contradiction since it violates the positive disutility of the labor supply.

[32] Observe that $D_x[\Upsilon(x) - \Upsilon'(x)x] = -\Upsilon''(x)x > 0$.

Given that the social planner sets $q_{i,t} = 1$ for all firms, it implies that firms' productions will differ under the efficient allocation only through different numbers of customers. To determine the optimal level of production, we only need to consider the FOC with respect to $l_{i,p,t}$:

$$\alpha \varphi_{c,i,t} z_{i,t} l_{i,p,t}^{\alpha-1} = \varphi_{L,t}.$$

Dividing this equation by the first order condition for $m_{i,t}$ in Equation (D.12), we get

$$z_{i,t} l_{i,p,t}^{\alpha-1} = \frac{C_t}{\alpha \Upsilon'(1)} \frac{\varphi_{L,t}}{\varphi_{\Upsilon,t}}.$$

Solving for $l_{i,p,t}$ from this equation and replacing it Equation (D.4) we have

$$\int_{j \in m_{i,t}} c_{i,j,t} dj = m_{i,t} C_t = z_{i,t} \left(\frac{C_t}{z_{i,t} \alpha \Upsilon'(1)} \frac{\varphi_{L,t}}{\varphi_{\Upsilon,t}} \right)^{\frac{\alpha}{\alpha-1}} \Rightarrow m_{i,t} = \left(\frac{z_{i,t}}{C_t} \right)^{\frac{1}{1-\alpha}} \left(\frac{\varphi_{L,t}}{\alpha \Upsilon'(1) \varphi_{\Upsilon,t}} \right)^{\frac{\alpha}{\alpha-1}}.$$

Finally, imposing the market clearing condition for matches in Equation (D.7) we get

$$m_{i,t} = \frac{z_{i,t}^{\frac{1}{1-\alpha}}}{\int_{i \in N_t} z_{i,t}^{\frac{1}{1-\alpha}} di}, \quad \forall i \in N_t.$$

□

D.8 Proposition 4

Proof. A few observations are useful in proving this Proposition. First, Lemma 2 allows us to ignore constraints in Equation (D.9) and Equation (D.10), solve for the optimal allocation of matches and use the results of Lemma 2 to characterize the allocation of marketing labor that satisfy these constraints. Second, since the planner's strategy in Lemma 2 is to always fully depreciate matches at the beginning of every period, the planner's entry and exit decision for a firm does not depend on $m_{i,t-1}$ and it is not relevant for the planner's decisions. Hence, the only distribution that we need to keep track of over time for the efficient allocation is the distribution of $z_{i,t}$.

Next, to characterize this distribution, define $n_t(z)$ as the density of operating firms at time t that have productivity z. Now consider an entry/exit policy for the planner at time t denoted by $\mathbf{1}_t(z)$ which takes the value of 1 if the planner pays the overhead cost of a firm with productivity z at time t—hence allowing the firm to be an operating firm at t. Then, $n_t(z)$ is given by:

$$n_t(z) = \mathbf{1}_t(z) \left[\lambda f^e(z) + \nu \int_{z_{-1} \in \mathbb{R}_+} f(z|z_{-1}) n_{t-1}(z_{-1}) dz_{-1} \right]. \tag{D.15}$$

Here, $f(z|z_{-1})$ denotes the conditional density of $z|z_{-1}$, which is governed by an AR(1) process per Equation (3.7) and $f^e(.)$ is the PDF of the productivity distribution of the entrants

from Equation (3.6). Equation (D.15) merely shows that the density of firms at time t with productivity z comes either from entrants with productivity z or operating firms that survived their exogenous exit shock and then transitioned to z. The planner can only decide whether it wants to keep these firms or not, but conditional on keeping them at time t their density is determined exogenously by the law of motion for the productivities.

The final step is to derive the aggregate production function of this economy using the results in Proposition 3. Using a similar approach for deriving Equation (3.19), we know that

$$\begin{aligned} L_{p,t} &\equiv \int_{i \in N_t} l_{i,p,t} di \\ &= C_t^{\alpha^{-1}} \int_{i \in N_t} \left(\frac{m_{i,t}^* q_{i,t}^*}{z_{i,t}} \right)^{\alpha^{-1}} di \\ &= C_t^{\alpha^{-1}} \left(\int_{z \in \mathbb{R}_+} z^{\frac{1}{1-\alpha}} n_t(z) dz \right)^{\frac{\alpha-1}{\alpha}}, \end{aligned}$$

where the first equation is the definition of aggregate labor allocated towards production, the second equation uses firm i's production function and the third equation plugs in the optimal allocation of q and m from Proposition 3. Hence, the aggregate production function of this economy can be written as:

$$C_t = \left[\int_{z \in \mathbb{R}_+} z^{\frac{1}{1-\alpha}} n_t(z) dz \right]^{1-\alpha} L_{p,t}^{\alpha}.$$

Given these observations about the planner's problem, and plugging in the results from Lemma 2 and Proposition 3, we can re-write the planner's problem as choosing C_t, L_t, $L_{p,t}$ and an entry/exit policy to maximize the life-time utility of the household subject to the (1) aggregate production function, (2) aggregate labor supply condition, and (3) law of motion for the distribution of productivity. Formally, the planner's revised problem is

$$\max_{\{C_t, L_t, L_{p,t}, \mathbf{1}_t(z)_{z \in \mathbb{R}_+}\}_{t \geq 0}} \sum_{t=0}^{\infty} \beta^t \left[\frac{C_t^{1-\gamma}}{1-\gamma} - \xi \frac{L_t^{1+\psi}}{1+\psi} \right] \tag{D.16}$$

$$s.t. \quad C_t = \left[\int_{z \in \mathbb{R}_+} z^{\frac{1}{1-\alpha}} n_t(z) dz \right]^{1-\alpha} L_{p,t}^{\alpha}$$

$$L_t = \chi \int_{z \in \mathbb{R}_+} n_t(z) dz + L_{p,t} + \bar{L}_{s,t} \tag{D.17}$$

$$n_t(z) = \mathbf{1}_t(z) \left[\lambda f^e(z) + \nu \int_{z_{-1} \in \mathbb{R}_+} f(z|z_{-1}) n_{t-1}(z_{-1}) dz_{-1} \right], \forall z \in \mathbb{R}_+ \tag{D.18}$$

$$n_{-1}(z) \text{ given.}$$

Next, for any t let $\beta^t \varphi_{C,t}$, $\beta^t \varphi_{L,t}$ and $\beta^t \varphi_{n,t}(z) dz$ be the Lagrange multipliers on the constraints in Equation (D.16), Equation (D.17) and Equation (D.18), respectively. The implied

first order conditions for C_t, L_t and $L_{p,t}$ are

$$C_t^{*-\gamma} = \varphi_{C,t}, \quad \xi L_t^{*\psi} = \varphi_{L,t}, \quad \alpha \frac{C_t^*}{L_{p,t}^*} \varphi_{C,t} = \varphi_{L,t}.$$

Combining these first order conditions, we arrive at the condition that the planner sets the marginal rate of substitution between consumption and leisure equal to the marginal product of labor (which we define as the wage in the decentralized version of this economy):

$$W_t^* \equiv \underbrace{\frac{\xi L_t^{*\psi}}{C_t^{*-\gamma}}}_{\text{MRS}} = \underbrace{\alpha \frac{C_t^*}{L_{p,t}^*}}_{\text{MPN}}.$$

Finally, since the entry and exit decision is a discrete choice, we have to compare the shadow cost/benefit of keeping a productivity type—setting $\mathbf{1}_{i,t}(z) = 1$—with the shadow cost/benefit of letting them exit—setting $\mathbf{1}_{i,t}(z) = 0$. Note that conditional on keeping a type, the FOC with respect to $n_t(z) > 0$ is

$$\varphi_{n,t}(z) = (1-\alpha) m_t^*(z) C_t^* \varphi_{C,t} - \chi \varphi_{L,t} dz + \beta \nu \int_{z' \in \mathbb{R}_+} \mathbf{1}_{t+1}(z') f(z'|z) \varphi_{n,t+1}(z') dz',$$

where $m_t^*(z) \equiv \frac{z^{\frac{1}{1-\alpha}}}{\int_{z \in \mathbb{R}_+} z^{\frac{1}{1-\alpha}} n_t(z) dz}$ is the number of customers that a firm with productivity z will get conditional on being kept in the economy as derived in Proposition 3. Dividing this equation by $\varphi_{C,t}$, replacing $\alpha C_t^* = W_t^* L_{p,t}^*$ and plugging in $W_t^* = \varphi_{L,t}/\varphi_{C,t}$ and $\varphi_{C,t} = C_t^{*-\gamma}$ we arrive at:

$$\frac{\varphi_{n,t}(z)}{\varphi_{C,t}} = m_t^*(z) C_t^* - m_t^*(z) W_t^* L_{p,t}^* - W_t^* \chi + \beta \nu \left(\frac{C_{t+1}^*}{C_t^*} \right)^{-\gamma} \int_{z' \in \mathbb{R}_+} f(z'|z) \mathbf{1}_{t+1}(z') \frac{\varphi_{n,t+1}(z')}{\varphi_{C,t+1}} dz',$$

where the right hand side characterizes the marginal benefit of keeping the firm, in units of consumption and the left hand side is the shadow cost of keeping the firm. Note that the planner can always set this to zero by making the firm exit the economy—meaning that this benefit/cost is bounded below by zero. Hence, the social value of the firm, G_{t^*}, can be defined as $G_t^*(z) \equiv \max_{\mathbf{1}_t(z)} \frac{\varphi_{n,t}(z)}{\varphi_{C,t}} \mathbf{1}_t(z)$ and be written recursively as:

$$G_t^*(z) = \max_{\mathbf{1}_t(z)} \mathbf{1}_t(z) \left\{ m_t^*(z) C_t^* - m_t^*(z) W_t^* L_{p,t}^* - W_t^* \chi + \beta \nu \left(\frac{C_{t+1}^*}{C_t^*} \right)^{-\gamma} \mathbb{E}[G_{t+1}^*(z')|z] \right\}.$$

Iterating this condition forward, and re-writing this in sequential form for firm i gives us the expression in the proposition. □

D.9 Proposition 5

Proof. Let $(C_t, L_{p,t}, L_{s,t}, L_t, N_t)_{t\geq 0}$ denote the equilibrium allocation. A log-linearization of $U(C_t, L_t) = \frac{C_t^{1-\gamma}}{1-\gamma} - \xi\frac{L_t^{1+\psi}}{1+\psi}$ around this allocation gives:

$$\Delta U(C_t, L_t) = C_t^{1-\gamma}\Delta\ln(C_t) - \xi L_t^{\psi} L_{p,t}(\Delta\ln(L_{p,t}) + \frac{L_{s,t}}{L_{p,t}}\Delta\ln(L_{s,t}) + \chi\frac{N_t}{L_{p,t}}\Delta\ln(N_t)) + \mathcal{O}(\|.\|^2).$$

Next, divide by $U_{c,t}C_t = C_t^{1-\gamma}$ and use the household's optimal labor supply condition $\xi\frac{L_t^{\psi}}{C_t^{-\gamma}} = W_t$ to get

$$\frac{\Delta U(C_t, L_t)}{U_{c,t}C_t} = \Delta\ln(C_t) - \frac{W_t L_{p,t}}{C_t}(\Delta\ln(L_{p,t}) + \frac{L_{s,t}}{L_{p,t}}\Delta\ln(L_{s,t}) + \chi\frac{N_t}{L_{p,t}}\Delta\ln(N_t)) + \mathcal{O}(\|.\|^2).$$

Finally, using the aggregate production function in Equation (3.18), replace $\Delta\ln(C_t) = \Delta\ln(Z_t) + \alpha\Delta\ln(L_{p,t})$, and, using the definition of the aggregate markup in Equation (3.20), replace the labor share in terms of the cost-weighted markup ($\frac{W_t L_{p,t}}{C_t} = \frac{\alpha}{\mathcal{M}_t}$) to get

$$\frac{\Delta U(C_t, L_t)}{U_{c,t}C_t} \approx \Delta\ln(Z_t) + \alpha(1-\mathcal{M}_t^{-1})\Delta\ln(L_{p,t}) - \alpha\mathcal{M}_t^{-1}(\frac{L_{s,t}}{L_{p,t}}\Delta\ln(L_{s,t}) + \chi\frac{N_t}{L_{p,t}}\Delta\ln(N_t)).$$

□

D.10 Proposition 6

Proof. Recall that from Equation (3.12) the relationship between a firm's production labor share and markup is given by:

$$\frac{W_t l_{i,p,t}}{p_{i,t}y_{i,t}} = \frac{\alpha}{\mu_{i,t}}.$$

Moreover, assuming $\delta = 1$, we can use the characterization of the firm's optimal marketing strategy in Equation (3.16) to write their marketing labor share of an operating firm as

$$\phi^{-1}\frac{W_t l_{i,s,t}}{m_{i,t}} = (p_{i,t} - mc_{i,t})q_{i,t}C_t \Leftrightarrow \frac{W_t l_{i,s,t}}{p_{i,t}y_{i,t}} = \phi(1-\mu_{i,t}^{-1}).$$

Combining these two equations we get that

$$W_t l_{i,s,t} = \phi p_{i,t}y_{i,t} - \phi\alpha^{-1}W_t l_{i,p,t}.$$

Finally, notice that

$$SGA_{i,t} \equiv W_t\chi + W_t l_{i,s,t} = \underbrace{SGAF_t}_{=W_t\chi} + \phi\underbrace{Sales_{i,t}}_{=p_{i,t}y_{i,t}} - \frac{\phi}{\alpha}\underbrace{COGS_{i,t}}_{=W_t l_{i,p,t}}.$$

□

D.11 Proposition 7

Proof. This result can be derived from combining Equations (D.10) and (D.10):

$$\frac{W_t(l_{i,p,t}+l_{i,s,t})}{p_{i,t}y_{i,t}} = \alpha\mu_{i,t}^{-1} + \phi(1-\mu_{i,t}^{-1}).$$

Notice that this is strictly decreasing in $\mu_{i,t}$ if and only if $\alpha > \phi$. Hence, the firm's revenue productivity of labor, the inverse of the equation above, is increasing in $\mu_{i,t}$ if and only if $\alpha > \phi$. □

E Computational Appendix

In this section, we present the details of the computation algorithm that solves and calibrates the model. We first describe the recursive representation of the firm's problem. Then, we describe the law of motion of firms and characterize the stationary distribution. Next, we describe the algorithm that solves the model, and the algorithm used in the calibration.

E.1 Solution Method

Firm's Recursive Problem In period t, a firm that decided to operate with a customer base of m_{-1} and productivity z solves the following dynamic programming problem (which corresponds to the sequential representation in Equation (3.8)):

$$\begin{aligned} & v_t(m_{-1},z) \equiv \max_{l_s,l_p,p}\left\{py - W_t l_p - W_t(l_s+\chi) + \beta\nu\frac{U_{c,t+1}}{U_{c,t}}\mathbb{E}\left[V_{t+1}(m,z')|z\right]\right\} \\ s.t. \quad & q = \left[1-\eta\ln\left(\frac{p}{D_t(1-\sigma^{-1})}\right)\right]^{\frac{\sigma}{\eta}} \\ & y = mqC_t = zl_p^{\alpha} \\ & m = (1-\delta)m_{-1} + \frac{l_s^{\phi}}{P_{m,t}}, \end{aligned}$$

where $V_t(m_{-1},z) \equiv \max\{0, v_t(m_{-1},z)\}$ denotes the endogenous exit choice.

Stationary Distribution Let $\mathcal{N}_t : M \times Z \to [0,1]$ denote the cdf of incumbent firms measured after the realization of idiosyncratic productivity shocks, but before exit decisions are made. The law of motion of the distribution of firms is given by:

$$\begin{aligned} \mathcal{N}_t(m,z') = & \int_{M\times Z} F(z'|z)\mathbf{1}_{\{m_t^*(m_{-1},z)\le m\}}\nu\mathbf{1}_{\{v_t(m_{-1},z)\ge 0\}}d\mathcal{N}_{t-1}(m_{-1},z) \\ & + \lambda\int_{M\times Z} F(z'|z)\mathbf{1}_{\{m_t^*(0,z)\le m\}}\mathbf{1}_{\{v_t(0,z)\ge 0\}}dF^e(z), \end{aligned}$$

where $F(z'|z)$ is the Markov chain given by the AR(1) productivity process, $F^e(z)$ is the productivity distribution of potential entrants and $m^*(m_{-1}, z)$ denotes the optimal policy for customer acquisition.

Solution Algorithm In steady state consumption is constant, so that $U_{c,t+1}/U_{c,t} = 1$. The algorithm for the numerical solution of the steady state of the model is as follows:

Step 0: Set up a grid for firm's state $S = M \times Z$. We choose 15 collocation points in each dimension. For Z, we use the 0.0001 and 0.9999 percentiles of the ergodic distribution of the AR(1) productivity process as the grid bounds. For M, the lower bound of the grid is 0 and the upper bound is chosen so that the largest customer base in the solution of any version of the model is smaller than the bound. Given these bounds, we construct power grids to concentrate grid points at lower values for m_{-1} and z.

Step 1: Guess values for C, $\tilde{W} \equiv W/(CD)$ and P_m.

Step 2: Solve firm's problem given $(C, \tilde{W}, P_m)$ by scaling the value function by $1/(CD)$ (this reduces the number of aggregate variables we need to solve for by one). We solve this problem by using projection methods to approximate both the value function $v_t(m_{-1}, z)$ and its expected value $\mathbb{E}\left[V_{t+1}(m, z')|z\right]$. We approximate these functions with the tensor product of a linear spline in the z dimension and a cubic spline in the m_{-1} dimension. We follow a two-step procedure to compute optimal policies. First, for a given candidate m, we compute q, l_s and l_p by solving the nonlinear FOC for q and using the production function and the law of motion of matches. Second, to optimize the value function with respect to m, we use the golden search method. Having approximated these values and guessed a vector of spline's coefficients, we combine an iteration procedure and a Newton solver to find the coefficient of the basis function. To compute the expectation in $\mathbb{E}\left[V_{t+1}(m, z')|z\right]$, we rely on the following approximation

$$\mathbb{E}\left[V_t(m, z')|z\right] = \sum_{i=1}^{50} \omega_i V_t\left(m, \exp(\rho \ln z + \varepsilon_i)\right).$$

To construct the nodes ε_i, we generate an equi-distant grid of 50 points from 0.0001 to 0.9999 and invert the cdf of the $\mathcal{N}(0, \sigma_z^2)$ distribution. To construct the weights ω_i, we discretize the normal distribution with a histogram centered around the nodes.

Step 3: To approximate the ergodic distribution of firms, we construct a finer grid with 100 and 500 points in the m_{-1} and z direction, respectively. Then, we solve the firm's problem once on the new grid using the approximation to the value functions from the previous step.

To find the ergodic distribution, we rely on the non-stochastic simulation approach by Young (2010). This method approximates the distribution of firms on a histogram based on the finer grid. Since both optimal policies and productivity shocks are allowed to vary continuously, we assign values of m and z that do not fall on points

in the grid in the following way. Let $s \equiv (m_{-1}, z)$ denote a firm's state. Then, the transition matrix for a firm's customer base can be constructed as:

$$Q_M(s, m'(s)) = \left[\mathbf{1}_{m'(s)\in[m_{j-1},m_j]} \frac{m'(s) - m_j}{m_j - m_{j-1}} + \mathbf{1}_{m'(s)\in[m_j,m_{j+1}]} \frac{m_{j+1} - y'(s)}{m_{j+1} - m_j} \right] \quad \text{(E.1)}$$

for all states s in the grid. That is, the transition matrix allocates firms in the histogram based on the proximity of the optimal policy to each point in the finer grid. The transition matrix for productivity shocks is approximated as $Q_Z = \sum_{i=1}^{200} \omega_i Q_{z,i}$, where $Q_{z,i}$ is similarly constructed as in Equation (E.1) for $z'(s) = \exp(\rho \ln z + \varepsilon_i)$. The overall transition matrix is then given by $Q = Q_Z \otimes Q_M$. Finally, the distribution of firms is obtained by iterating until convergence the approximation to the law of motion

$$\boldsymbol{\mathcal{N}} = Q' \left(\nu \mathbf{1}_{v(\boldsymbol{s})\geq 0} \boldsymbol{\mathcal{N}} + \lambda \mathbf{1}_{v(\boldsymbol{s})\geq 0} F^e \right),$$

where F^e is an approximation of the distribution of entrants on the finer grid.

Step 4: Compute aggregate variable X from firms' vectorized policies $x(\boldsymbol{s})$ as $X = (\nu \mathbf{1}_{v(\boldsymbol{s})\geq 0} \boldsymbol{\mathcal{N}} + \lambda \mathbf{1}_{v(\boldsymbol{s})\geq 0} F^e)' x(\boldsymbol{s})$. Compute the residual vector

$$1 = \int_0^N m_i di, \quad 1 = \int_0^N m_i \Upsilon(q_i) di, \text{ and } \quad 1 = \frac{W}{\xi C^\gamma L^\psi}.$$

If the distance is small, stop. Otherwise, update $(C, \tilde{W}, P_m)$ with a Newton method and go to **Step 2**.

E.2 Estimation routine

We estimate the parameters of the model via the Simulated Method of Moments (SMM). More specifically, we choose a set of parameters $\mathcal{P}$ that minimizes the SMM objective function

$$\left(\frac{\boldsymbol{m}_m(\mathcal{P})}{\boldsymbol{m}_d} - 1 \right)' \boldsymbol{W} \left(\frac{\boldsymbol{m}_m(\mathcal{P})}{\boldsymbol{m}_d} - 1 \right),$$

where $\boldsymbol{m}_m$ and $\boldsymbol{m}_d$ are a vector of model simulated moments and data moments, respectively, and $\boldsymbol{W}$ is a diagonal matrix. To compute the model simulated moments we follow these steps:

Step 1: Given a vector of parameters $\mathcal{P}$, we find the steady state of model. For this, we slightly modify the previous algorithm. Since in the estimation we normalize aggregate output $C = Y = 1$ and the normalized wage $\tilde{W} = 1$ (see Step 1 of the solution algorithm) with the free parameters (λ, ξ), we need to solve for only one aggregate variable, P_m.

Step 2: Simulate 100,000 firms for 150 periods and compute model moments using data from the last 25 periods. When matching moments based on the entire US economy, we use data from all simulated firms. When matching moments based on Compustat data, we impose a filter that mimics selection into Compustat based on

firm age and size. On the age dimension, we restrict the simulated sample to those firms that are at least 7 years old, as in Ottonello and Winberry (2018). On the size dimension, we restricted the sample to firms with sales above 19% of the average sales in the simulated economy. This cutoff corresponds to the ratio of the 5th percentile of the sales distribution in Compustat (USD1.06 million) to the average firm sales in SUSB (USD5.7 million) in 2012.

To minimize the SMM objective function and have confidence of reaching the global minimum, we follow a two-step procedure in the spirit of Arnoud, Guvenen and Kleineberg (2019). In the first step, we construct 500 quasi-random vectors of parameters $\mathcal{P}$ from a Halton sequence, which is a deterministic sequence designed to evenly cover the parameter space. After computing the SMM objective in those points, we choose the 30 parameters vectors with the lowest objective values. In the second step, we initiate a local Nelder-Mead optimizer from each of the 30 starting points and select the local minimum with the lowest objective value.

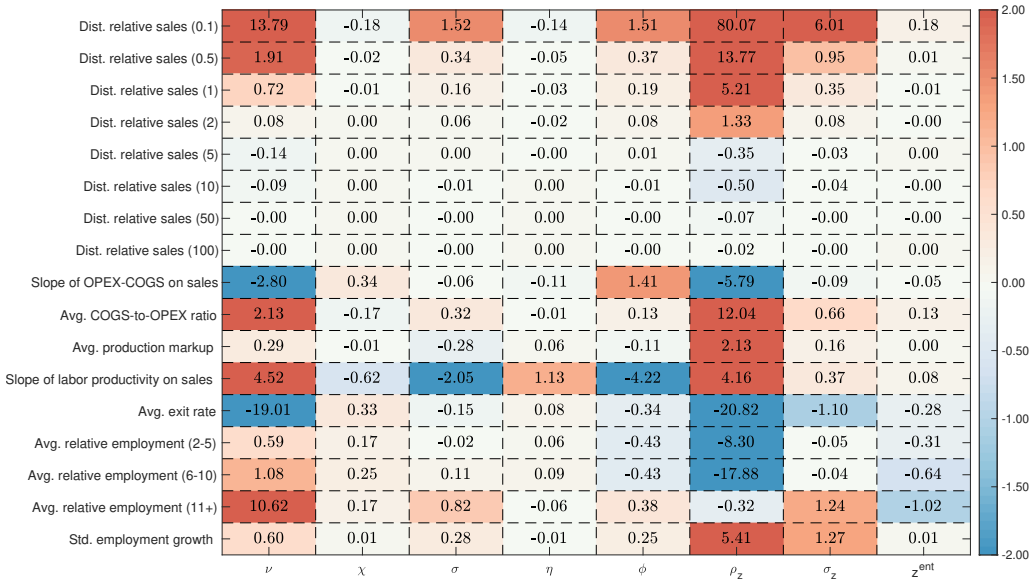
F Additional Model Analysis

F.1 Identification of Model Parameters

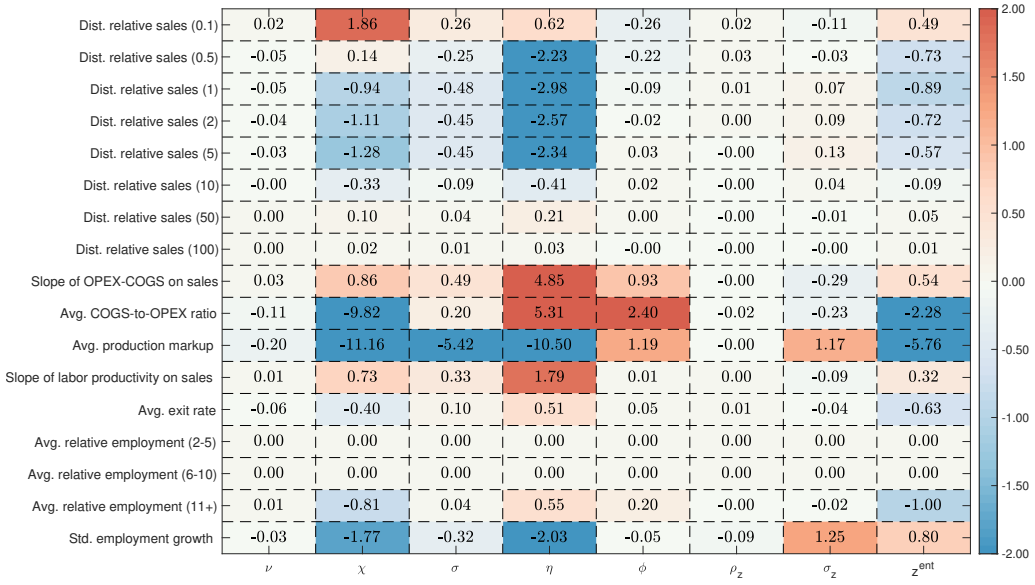
In this Section, we formally guide the discussion of the identification of model parameters. Panel A of Figure F.1 shows the *local* elasticity of simulated moments (rows) with respect to parameters (columns), evaluated at the calibrated parameters. In general, the intuition behind the choice of targets is borne out in the model. For example, a higher exogenous exit rate $1-\nu$ mechanically increases the average exit rate. Similarly, a higher super-elasticity of demand η increases the co-movement between revenue labor productivity and sales. A higher elasticity of the matching function ϕ makes the positive relationship between a firm's SGA expenses and sales stronger and reduces the co-movement between labor productivity on sales, as predicted in the simple version of the model in Propositions 6 and 7. The persistence of productivity shocks ρ_z affects multiple moments, but it affects most strongly the dispersion of the sales distribution. Finally, a smaller average productivity of entrants $\bar{z}_{ent}$ increases the relative size of old firms.

We complement this discussion by analyzing the sensitivity measure developed by Andrews et al. (2017), which show the sensitivity of model parameters with respect to targeted moments. To make the numbers more comparable, we convert this measure into elasticities and plot $\left(J'(\mathcal{P})WJ(\mathcal{P})\right)^{-1}J'(\mathcal{P})Wm_m(\mathcal{P})/\mathcal{P}$, where $J(\mathcal{P})$ is the Jacobian evaluated at calibrated parameters, W is the weighting matrix and $m_m(\mathcal{P})$ are the model moments evaluated at calibrated parameters. Panel B of Figure F.1 shows that the overhead cost χ is quite sensitive to the average COGS-to-OPEX ratio in the data. Similarly, the elasticity of substitution σ is sensitive to the average production markup and the standard deviation of the productivity shock σ_z is most influenced by the standard deviation of employment growth.

Figure F.1: Parameter Identification



(a) Sensitivity of Moments



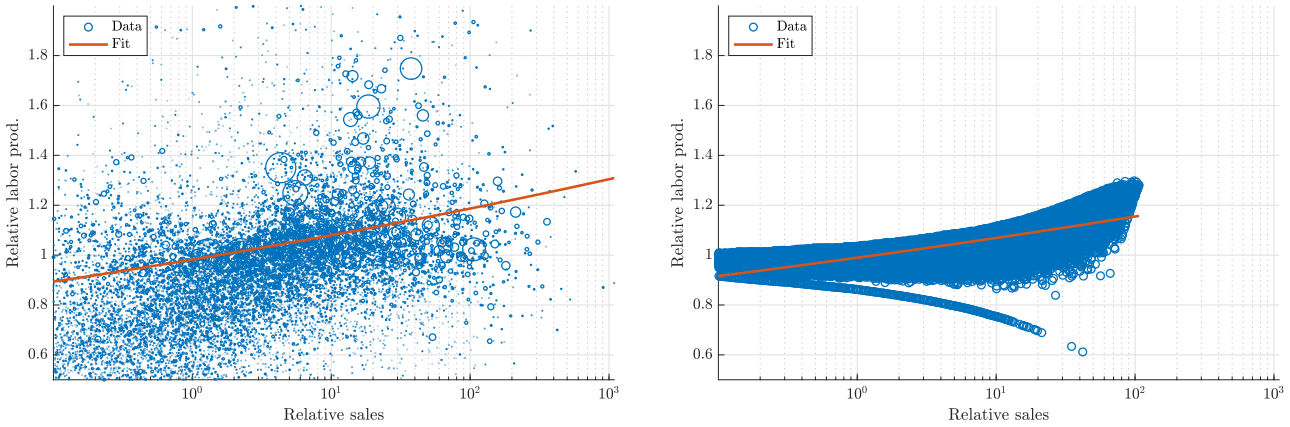
(b) Sensitivity of Parameters

Notes: Panel A shows the sensitivity of simulated moments to parameters by computing the local elasticity of moments with respect to parameters. Panel B shows the sensitivity of calibrated parameters to moments by constructing the sensitivity measure of Andrews et al. (2017) and converting it into an elasticity. Both measures are evaluated at the calibrated parameters.

F.2 Additional Calibration Figures

Here we provide additional results about the goodness of fit of the calibrated model. Figure F.2 plots the relationship between relative revenue productivity of labor and relative sales in the data (SUSB) and the model (both the raw data and a linear fit). The model is able to match the positive association between these variables well. Figure F.3 shows the relationship between relative SGA and relative sales in the data (Compustat) and the model. Although the "Compustat-equivalent" sample from the model does not generate all the dispersion in relative sales as in the data, the relationship with relative SGA is well matched in the overlapping range of relative sales. Finally, Figure F.4 plots the average COGS-to-OPEX ratio as a function of firm's age and size in the data (Compustat) and the model. Here, age is normalized as time since entry into Compustat (which in the model occurs after the 7th year). In the model, the composition of firms' costs exhibits a strong size profile and a weak age profile, as in the data.

Figure F.2: Model Fit: Labor Productivity and Sales



(a) Data (b) Model

Notes: This figure plots the relationship between relative revenue productivity of labor and relative sales. Panel A and B show the relationship obtained from the SUSB data and the model simulated data, respectively. Each figure includes the best linear fit of the data. The x-axis is in log scale.

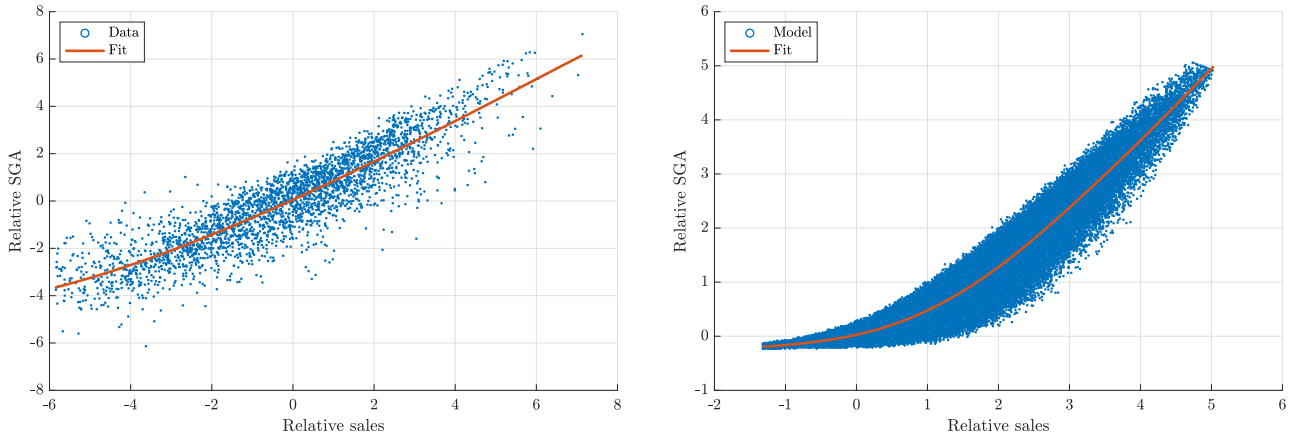
F.3 Additional Analysis of Model in Steady State

In this section, we further describe how the model works in steady state. First, we describe firms' optimal policies. Then, we show the average firm dynamics, taking selection into account.

Firms' Optimal Policies Figure F.5 shows firms' steady state optimal policy functions for three productivity levels (the 25th, 50th and 75th percentile of the marginal productivity distribution in steady state). The y-axis on the right plots the marginal distribution of the relative customer base.

While optimal spending in $l_{i,s,t}$ decreases with the size of the customer base, production labor is increasing in a firm's customer base. Thus, when firms have a small customer base,

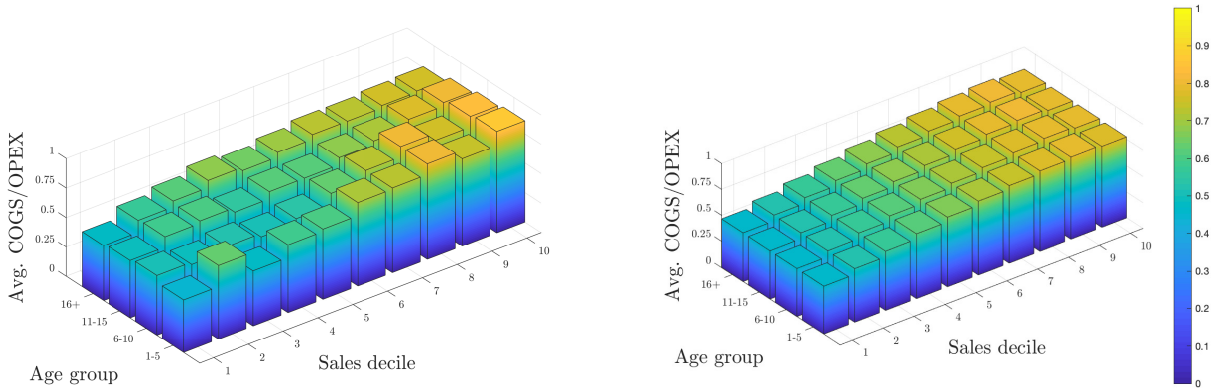
Figure F.3: Model Fit: SGA and Sales



(a) Data (b) Model

Notes: This figure plots the relationship between relative spending in SGA and relative sales. Panel A and B show the relationship obtained from the Compustat data and the model simulated data, respectively. The model data is obtained by simulating the model and restricting the sample to firms that are at least 7 years old and have sales above 19% of the average sales in the simulated economy (which corresponds to the ratio of the 5th percentile of the sales distribution in Compustat to the average sales in SUSB). Each figure includes the local linear kernel best fit of the data.

Figure F.4: Steady-State COGS/OPEX by Size and Age



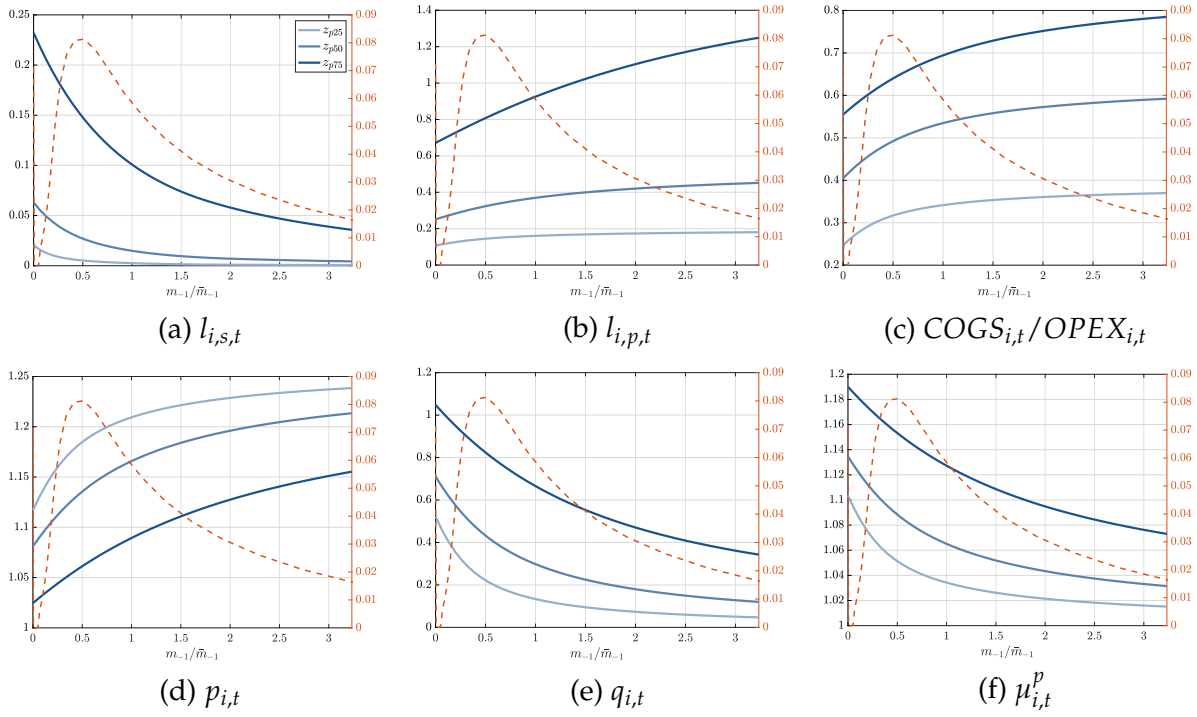
(a) Data (b) Model

Notes: This figure plots the average COGS-to-OPEX ratio as a function of a firm's age and size. Panel A and B show the relationship obtained from the Compustat data and the model simulated data, respectively. The model data is obtained by simulating the model and restricting the sample to firms that are at least 7 years old and have sales above 19% of the average sales in the simulated economy (which corresponds to the ratio of the 5th percentile of the sales distribution in Compustat to the average sales in SUSB). Age is normalized as years since entry into Compustat, which in the model corresponds to year 7.

they spend more resources to increase it. However, due to decreasing returns to customer accumulation, firms increase their customer base gradually over time. As firms grow, they spend less on customer acquisition and more on producing goods to satisfy the growing

demand. This is reflected in a firm's cost structure—the average COGS-to-OPEX ratio is also increasing in m_{-1}. As total output increases due to a larger customer base, the marginal cost of production also increases since production is subject to decreasing returns. This raises the price charged by the firm, which in turn reduces the consumption per capita $q_{i,t}$ and optimal markups. The Figure also shows that, for a given level of m_{-1}, spending in $l_{i,s,t}$ is increasing in a firm's productivity. A higher productivity allows firms to charge lower prices and higher markups. Thus, profits per marginal customer are increasing in productivity, which incentivizes firms to accumulate customers more quickly by spending more on $l_{i,s,t}$.

Figure F.5: Firms' Optimal Policies



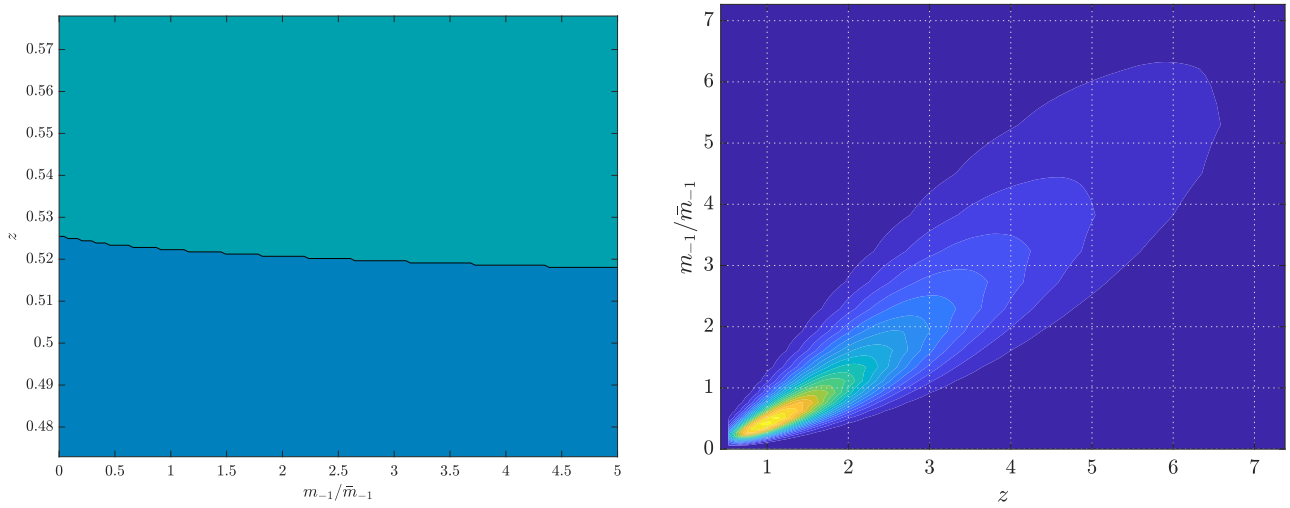
(a) $l_{i,s,t}$ (b) $l_{i,p,t}$ (c) $COGS_{i,t}/OPEX_{i,t}$

(d) $p_{i,t}$ (e) $q_{i,t}$ (f) $\mu^p_{i,t}$

Notes: These figures plot firms' policy functions in steady state. Each figure shows policies as a function of relative customer base for three levels of productivity: the 25th, 50th and 75th percentile of the stationary productivity distribution. The y-axis on the right plots the stationary marginal distribution of the relative customer base.

Figure F.6 plots firms' optimal exit policies and the stationary joint distribution of (m_{-1}, z). Panel A shows the threshold productivity $z^*(m_{-1})$ such that if $z < z^*(m_{-1})$, the firm optimally chooses to exit. The figure shows that $z'^*(m_{-1}) < 0$, that is, firms with larger customer base are able to survive large productivity shocks without the need to exit the market. Although a lower productivity reduces markups and profits per customer, aggregate profits are increasing in a firm's customer base. Panel B shows that in steady state there is a positive correlation between firms' productivities and customer bases: more productive firms have on average a larger customer base.

Figure F.6: Optimal Exit and Stationary Distribution



(a) Optimal exit policy

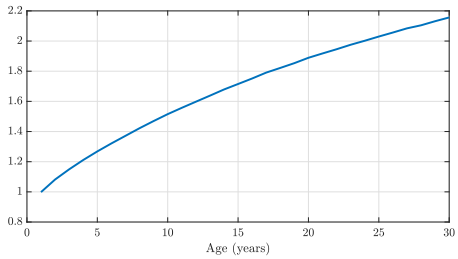
(b) Stationary distribution of (m_{-1}, z)

Notes: Panel A plots the exit threshold $z^*(m_{-1})$ such that if $z < z^*(m_{-1})$, the firm optimally chooses to exit. Panel B shows the contour plot of the stationary joint distribution of (m_1, z), censored at the 99th percentile of each variable.

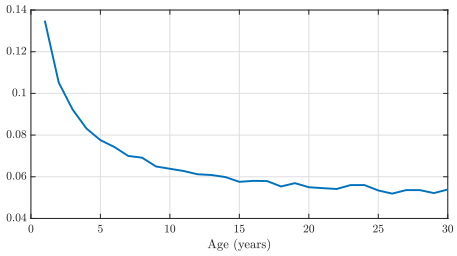
Average Firm Dynamics with Shocks Figure F.7 plots the average firm dynamics taking selection into account. To construct this figure, we simulate a cohort of firms that starts with a zero customer base and draws productivities from the distribution of entrants. As firms are subject to productivity shocks, some of them decide to exit over their lifetime. The figure plots the average of each variable across firms that survived up to a given age.

Firms start with lower productivity, which grows over time due to the calibrated lower productivity of entrants and endogenous exit. Conditional on a productivity level, firms front load spending on customer acquisition, and the average customer base and marginal costs rise rapidly for young firms. This in turn increases prices and reduces average output per customer and markups. Over time, only the most productive firms survive, so the average marginal cost and price declines and output per customer and markups increase.

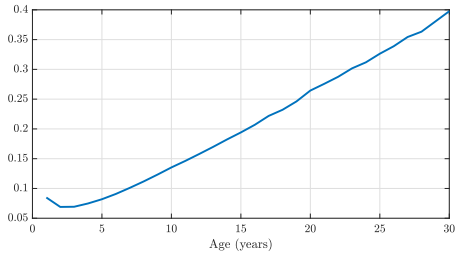
Figure F.7: Average Firm Dynamics



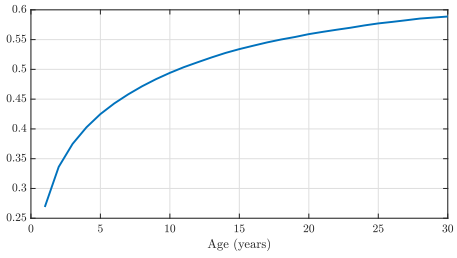
(a) Productivity



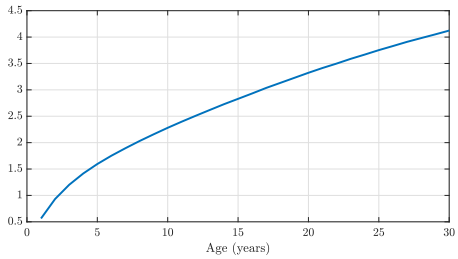
(b) Exit rate



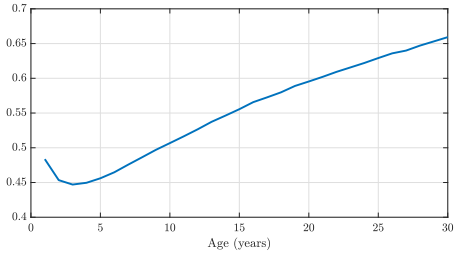
(c) $l_{i,s,t}$



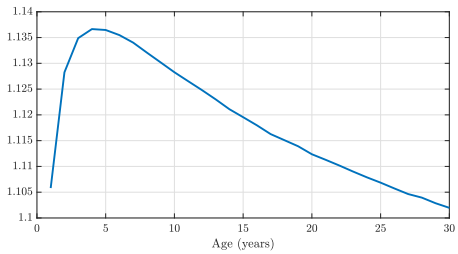
(d) $COGS_{i,t}/OPEX_{i,t}$



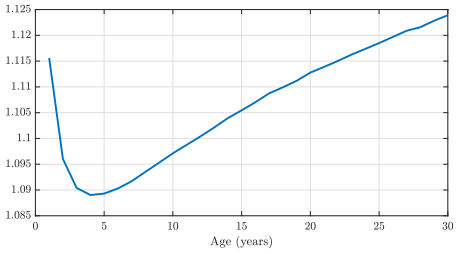
(e) $m_{i,t}$



(f) $q_{i,t}$



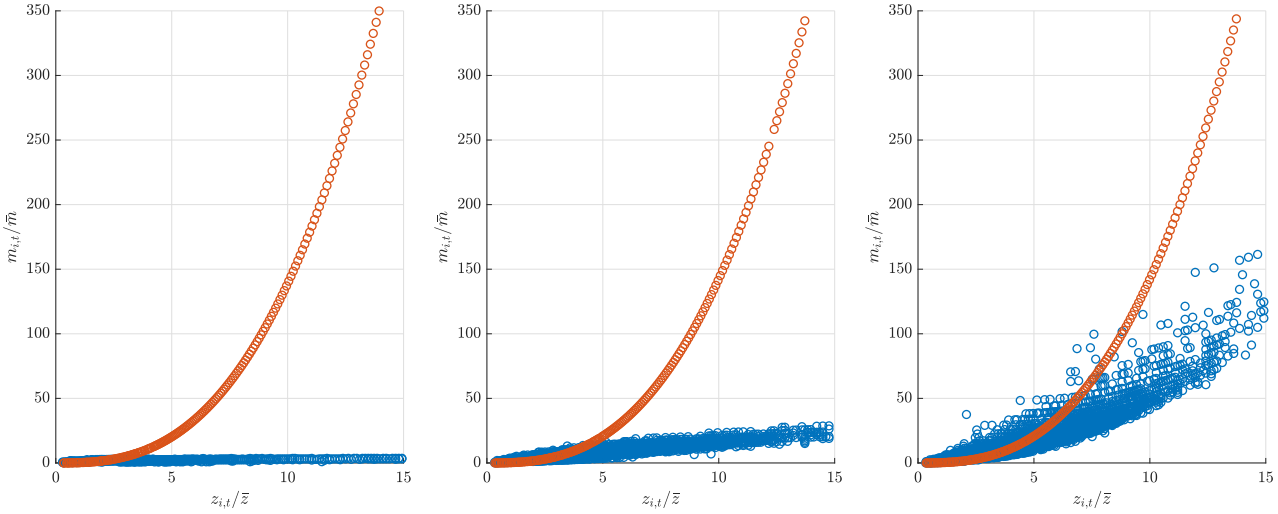
(g) $p_{i,t}$



(h) $\mu^p_{i,t}$

Notes: The figure plots the average firm dynamics, which are obtained by simulating a cohort of firms that start with $m_{-1} = 0$, draw z from the distribution of entrants, and experience productivity shocks over their lifetime. Each figure plots the average of a variable as a function of firms' age across firms that survived up to that age.

Figure F.8: Allocation of Customers: Equilibrium vs. Efficient Allocation



(a) Low ϕ (b) Baseline ϕ (c) High ϕ

Notes: This figure shows a scatter plot between relative productivity $z_{i,t}/\bar{z}$ and relative customer bases $m_{i,t}/\bar{m}$, for both the equilibrium and the social planner's allocation. We present three plots by varying the value of ϕ, while keeping the remaining parameters fixed at the values in the baseline calibration. Low ϕ corresponds to 0.25, baseline to 0.53, and high to 0.75.

G Calibration of the Restricted Model

Table G.1: Model Parameters: Restricted Model

Parameter	Description	Value
	Panel A: Fixed Parameters	
β	Annual discount factor	0.960
γ	Elast. of intertemporal substitution	2.000
ψ	Frisch elasticity	1.000
α	Decreasing returns to scale	0.640
δ	Prob. of losing customer	1.000
	Panel B: Calibrated Parameters	
χ	Overhead cost	0.704
σ	Avg. elasticity of substitution	5.998
η	Superelasticity	0.559
ν	Exog. survival probability	0.960
ρ_z	Persistence of productivity shock	0.981
σ_z	SD of productivity shock	0.259
$\bar{z}_{ent}$	Mean productivity of entrants	-2.379
λ	Mass of entrants	0.123
ξ	Disutility of labor supply	2.246

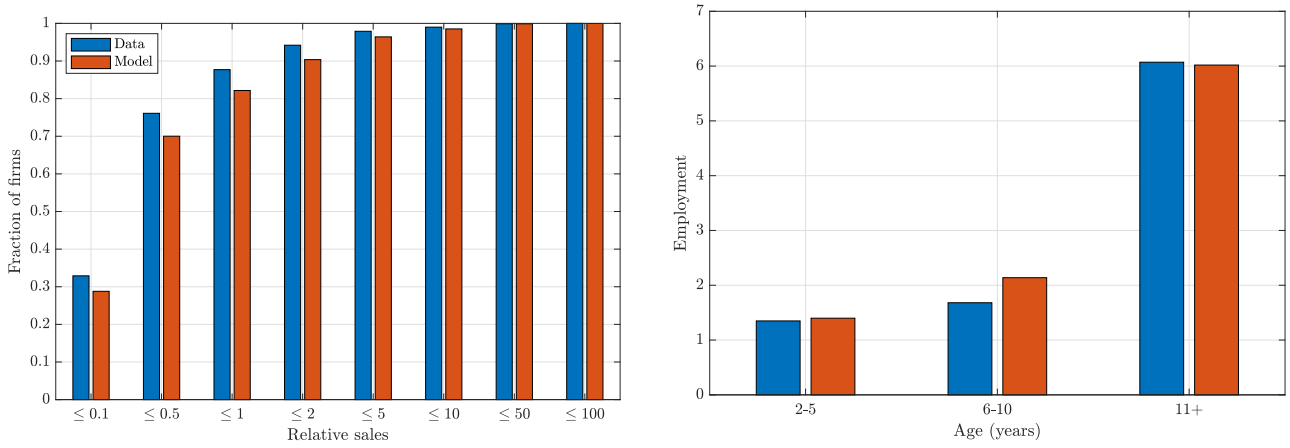
Notes: This table shows the calibration of the Restricted model with an exogenous customer Base. Panel A contains parameters externally chosen. Panel B contains parameters internally calibrated to match moments presented in Table G.2 and Figure G.1.

Table G.2: Targeted Moments: Restricted Model

Moment	Data	Model
Avg. COGS-to-OPEX ratio	0.660	0.667
Avg. cost-weighted production markup	1.250	1.266
Slope labor prod. on sales	0.036	0.035
Avg. exit rate	0.073	0.073
SD. employment growth	0.416	0.436

Notes: This table shows the set of moments targeted in the calibration of the Restricted model with an exogenous customer base. Avg. COGS-to-OPEX ratio refers to the average of the ratio across firms. Avg. cost-weighted production markup corresponds to the COGS-weighted average markup from Edmond et al. (2018). These moments were computed using data from Compustat in 2012. Slope of labor prod. on sales corresponds to the OLS coefficient of the sales-weighted regression of relative revenue labor productivity on relative sales from Edmond et al. (2018), restricting the sample to firms with relative sales above one. This moment was computed using data from the SUSB in 2012. The average exit rate was obtained from the BDS in 2012. The standard deviation of annual employment growth for continuing establishments is obtained from Elsby and Michaels (2013). Growth rate of variable x is computed as in Davis and Haltiwanger (1992): $(x_{i,t} - x_{i,t-1})/0.5(x_{i,t} + x_{i,t-1})$. The last column shows the model counterparts of each moment, which was obtained by simulating a panel of firms and computing each moment with the simulated data. In the model, we account for selection into Compustat by restricting the simulated sample of firms to those that are at least 7 years old and have sales above 19% of the average sales in the simulated economy (which corresponds to the ratio of the 5th percentile of the sales distribution in Compustat to the average sales in SUSB).

Figure G.1: Restricted Model Fit: Firm Size



(a) Distribution of Sales

(b) Avg. Size by Age

Notes: This figure shows moments targeted in the calibration of the Restricted model with an exogenous customer base. Panel (a) shows the model fit of the distribution of relative sales. The distribution of relative sales is obtained from the SUSB in 2012. Panel (b) shows the model fit of average employment, relative to 1 year-old firms, by firm age. Average firm employment by age group was obtained from BDS in 2012. In the calibration exercise, we only target the relative size of firms older than 10 years.

Figure G.2: Distribution of Markups: Baseline vs Restricted Model

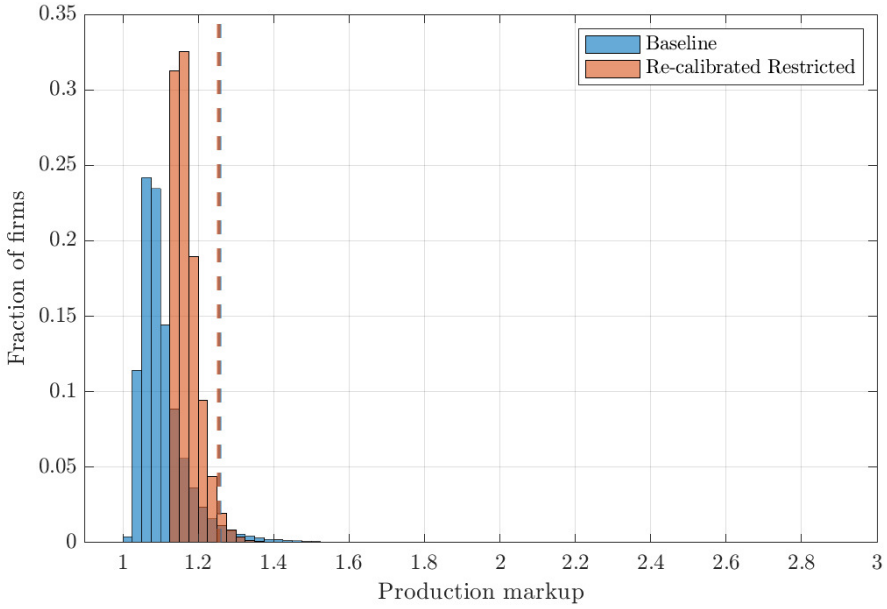


Figure G.3

Notes: The figure plots the distribution of production markups in the baseline and *re-calibrated* restricted models. The vertical dashed lines show the average cost-weighted production markup in each model.