CESifo WORKING PAPERS

8633
2020

October 2020

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

Hassan Afrouzi, Andres Drenik, Ryan Kim

CESifo

Impressum:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editor: Clemens Fuest
https://www.cesifo.org/en/wp

An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: https://www.cesifo.org/en/wp

CESifo *Working Paper No. 8633*

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

Abstract

How are a firm's size and market power related to one another? Combining micro-data about producers and consumers, we document that while firms mainly grow by selling to more customers, their markups are only associated with their average sales per customer. To study the macroeconomic implications of these facts, we develop a model of firm dynamics with endogenous customer acquisition and variable markups. Relative to a model without customer acquisition, our model generates higher concentration at the top, but a lower aggregate markup. Our quantitative analysis reveals large welfare and efficiency losses due to (mis)allocation of customers across firms. By increasing market concentration among the most productive firms, the efficient allocation achieves 11% higher aggregate productivity and 15% higher output.

JEL-Codes: D240, D420, D610, E220.

Keywords: customer acquisition, misallocation, concentration, markups.

Hassan Afrouzi
Columbia University / New York / USA
hassan.afrouzi@columbia.edu

Andres Drenik
Columbia University / New York / USA
ad3376@columbia.edu

Ryan Kim
Johns Hopkins University
Washington DC / USA
rkim59@jhu.edu

First Draft: August, 2020
This Draft: October, 2020
We thank Luigi Caloi for providing superb research assistance. We also thank John Asker, David Baqaee, Ariel Burstein, Matthieu Gomez, Virgiliu Midrigan, and participants at the VMACs Junior Conference, as well as Cleveland Fed and Columbia Macro Lunch seminars for valuable comments and suggestions. 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.

1 Introduction

Conventional macroeconomic models of endogenous markups generate a tight link between firm size and market power.[1] However, a potentially relevant distinction for market power is that firms can gain market share either because they *sell to more customers* (extensive margin of demand) or because they *sell more per customer* (intensive margin of demand). This poses two natural questions. First, which one of these two factors contributes more to the growth of firms' sales? Second, and more importantly, how does firms' market power depend on each of these margins of concentration? The answers to these questions have significant implications for understanding the relationship between market power and concentration.

In this paper, we make three contributions. First, we utilize a novel merged dataset between individual product-level consumption data from the ACNielsen Homescan Panel and producer-level data from Compustat to uncover the role of the intensive and extensive margins of demand for concentration and market power. Our empirical analysis reveals that while firms grow mostly through selling to more customers, their market power—measured by price-cost markups—is mainly associated with their average sales per customer. Second, we develop a theoretical framework with endogenous customer acquisition and variable markups in order to study the implications of these facts for (mis)allocation of customers across firms, concentration, market power, and welfare. Third, we devise an identification strategy to measure the magnitude of the distortions in allocation of customers across firms and quantify the efficiency losses from a novel source of misallocation: the misallocation of demand. Relative to the efficient allocation, we find that aggregate TFP and output are 10.8% and 14% lower in the equilibrium, respectively. A further decomposition of aggregate TFP shows that three-quarters of these losses are due to misallocation of demand.

Our analysis begins by answering our motivating questions. Merging the ACNielsen Homescan Panel and Compustat datasets, we document the following three facts. First, as firms enter a market and grow, much of their sales growth is explained by an expansion of their customer bases as opposed to an increase in their average sales per customer. Second, firms' non-production expenses are associated with the acquisition of new customers but not with their average sales per customer, nor the retention of their existing customers. Third, markups are not correlated with firms' number of customers. Instead, they are highly associated with firms' average sales per customer. These empirical facts serve as the foundation of the theory that we develop in order to understand the relationship between the

[1] This is true in a variety of models. See, for instance, Edmond, Midrigan and Xu (2018) for a model with Kimball demand; Atkeson and Burstein (2008) for a model with nested CES and oligopolistic competition; and Rotemberg and Woodford (1992) for a model with oligopolistic competition and implicit collusion.

different sources of firm size and market power.

In our micro-founded model, firms with heterogeneous productivity launch marketing campaigns to acquire new customers and face a semi-kinked demand schedule (à la Kimball, 1995) from each customer—which implies a higher demand elasticity at higher relative prices. 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 empirical facts: our model allows firms to grow through both extensive and intensive margins of demand (first fact), implies an endogenous relationship between sales and non-production costs (second fact), and creates a comovement between markups and sales per customer, but not necessarily with the number of customers (third fact).

One of the main implications of endogenous customer acquisition in our model is higher concentration but lower aggregate market power relative to a model without customer acquisition. The reason for this result is that in our model more productive firms grow mainly through the extensive margin of demand. Since this margin only shifts the firms' demand curves, our model generates higher concentration among the more productive firms without increasing their market power. It is important to note that this arises from firms' choices on how to grow. Akin to conventional models of endogenous markups, more productive firms in our model still have the option to grow by selling more per customer and have higher markups—which happens to some degree in the equilibrium. Nonetheless, once we endogenize firms' choices to grow through both margins, the extent of their growth through the intensive margin is mitigated by their desire to grow through the extensive margin. Compared to their counterparts in conventional models, such firms charge lower markups but make higher profits through their larger bases.

To further study the macroeconomic implications of endogenous customer acquisition, we also characterize the efficient allocation in our model. We show that under the efficient allocation, sales per customer are equalized across all firms but the distribution of sales is even more concentrated at the top of the productivity distribution. This is in contrast to the efficient allocation in conventional models in which demand is only modeled in the intensive margin. In those models, the planner can only achieve a higher aggregate productivity by assigning more demand per customer to more productive firms. But the gains from these types of allocations are small as they distort the allocation of relative demand of weakly substitutable varieties. In our model, however, the planner does not face this trade-off: she increases aggregate efficiency by allocating more customers towards more productive firms, while avoiding relative demand distortions by equalizing the relative demand per customer across all varieties.

In addition, even though our model implies a lower aggregate markup relative to conventional models, welfare losses are potentially larger. This 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 homogeneous allocation of customers is still feasible in our model, it is no longer exogenously imposed on the planner. Hence, despite a lower aggregate markup, our model predicts potentially larger welfare losses as long as the equilibrium allocation of customers is sufficiently distorted. This last observation unveils a novel source of *demand misallocation* that emerges in the presence of endogenous customer acquisition.

Next, we quantify the model in order to provide a measurement of the differences in the allocation of customers, aggregate productivity and welfare between the equilibrium and the efficient allocation. One of the key challenges posed by this analysis is the identification of model parameters that determines the equilibrium allocation of customers across firms. Based on the predictions of the model, we provide an identification strategy that can be implemented with available data on firms' sales and cost structures. At the core of this strategy is the contemporaneous comovement between a firm's sales and its non-production expenses (conditional on production expenses, for reasons made clear below), which is indicative 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 compare our equilibrium with those obtained in a restricted version of our model, in which each firm receives an exogenous number of customers. This restriction has large aggregate consequences. By forcing firms to switch their sales growth strategy, from expanding their customer base to increase 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 their customer base in 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 make predictions about changes in market power.

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 household under the efficient allocation is 13.6%, where 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. The planner is able to achieve 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%. To illustrate the role played by the (mis)allocation of customers, we reproduce the comparison between the efficient and equilibrium allocations for different degrees of decreasing returns to advertising. By moving half-way from the calibrated model to an economy with constant returns to advertising, differences become less pronounced: aggregate TFP is "only" 3.2% higher, welfare is 4% higher and the efficient degree of concentration is 15% higher.

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 concentration and its ability to exert market power.[2] Our contribution to this literature is twofold. From an empirical perspective, we disentangle sources of market power and document that the relevant source of market power is 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 model generates higher concentration *and* lower aggregate market power than conventional models. Thus, our model cautions against using measures of concentration to make predictions about changes in market power.[3]

Our model builds upon a large body of work that analyzes the consequences of endogenous customer acquisition through marketing or advertising activities (see, e.g., Arkolakis, 2010; Drozd and Nosal, 2012; Sedláček and Sterk, 2017; Perla, 2019), rather than prices.[4] This choice is guided by the fact that we find no relationship between markups and the size of the customer base after controlling for a firm's average sales per customer. This is also aligned with Fitzgerald, Haller and Yedid-Levi (2016) and Fitzgerald and Priolo (2018), who in different settings present model-based and empirical evidence in favor of models with customer acquisition based on marketing-like activities. In this sense, our paper is also related to Kaplan and Zoch (2020), who analyze the implications of labor-intensive expansionary activities of firms. Relative to these papers, we analyze a model with endogenous customer acquisition based on marketing activities and variable markups due to semi-kinked demand on the intensive margin. We embed this framework in a standard model of firm dynamics (as in Hopenhayn, 1992) to study the misallocation of demand across firms by providing a comparison of the equilibrium with the efficient allocation.

Our paper is also related to the literature that analyzes the role of misallocation of pro-

[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).

[3] 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, e.g., 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).

duction inputs across firms in affecting aggregate TFP (see the seminal work by Restuccia and Rogerson, 2008; Hsieh and Klenow, 2009). The literature has focused on multiple sources of misallocation: endogenous market power (Edmond et al., 2018; Peters, 2019), 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). Instead, we quantify the contribution of the misallocation of customers to aggregate productivity.

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 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. Section 7 discusses the assumptions of the model and the implications for measurement of markups. Finally, Section 8 concludes.

2 Empirical Analyses

To guide our empirical analyses, consider the following exact decomposition of log sales of firm i, $\ln S_i$:

$$\ln S_i = \ln m_i + \ln \left(p_i D(p_i)\right) \tag{2.1}$$

where m_i is the number of customers firm i faces, p_i is the price of product firm i sells, and $D(p_i)$ is the quantity purchased by each customer. Standard models in macro and international economics literature focus on the $p_i D(p_i)$ part of S_i and abstract away from differences in the size of the customer base, m_i, across firms.

With micro-level data, we document three new facts that shed light on the importance of m_i in understanding: (1) how firms grow over time, (2) how they distribute their expenses, and (3) how they charge their price-cost markups. To summarize, we find that:

Fact 1. As firms enter and grow, much of their sales growth arises from the acquisition of new customers (m_i) as opposed to the increase in sales per each customer ($p_i D(p_i)$).

Fact 2. The acquisition of new customers (m_i) is associated with the firm's non-production expenses, whereas sales per customer ($p_i D(p_i)$) are not.

Fact 3. Unlike standard models that link markups and market share, markups are only correlated with average sales per customer ($p_i D(p_i)$) and are unrelated to the number of customers (m_i).

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

2.1 Data Description

We build up a detailed customer-firm-matched dataset to decompose firms' sales into differences in the size of their customer base and differences in sales per customer, and analyze each margin related to firms' growth, cost structure, and markups.

To identify the number of customers each firm faces, we use the ACNielsen Homescan Panel, which was 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. A barcode, a unique universal product code (UPC) allocated to each product, is used to scan and store product information. Nielsen samples households and provides in-home scanners to make those sampled households record their purchases of products with barcodes. 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] The data assigns a broad product group for each product, such as pet food and school supplies, and record information on which retailer each household visited to purchase products at a given point of time. Nielsen claims that 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 period of 2004-2016.

Next we merge the Nielsen database with GS1 US Data Hub to group products by the producing firm and bring in other firm-level information. GS1 is the business entity that provides barcodes to firms. Their data record the firm name for each UPC available in Nielsen data and allow us to link customer information with its producer information. The definition of a firm is based on the unit that purchased barcodes from the GS1; firms in our data include both manufacturers and retailers.

We also make use of firm-level balance sheet information from Compustat to analyze firms' cost structures.[6] Compustat includes panel data on publicly traded firms since 1960.

[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 those that look suspicious.

With the caveat that it only covers publicly listed firms, it constitutes the main source of data for firm-level analysis in the US and has been used by 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 (SGA, non-production costs). According to Compustat, COGS is *"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 typical categories are the cost of labor and intermediate inputs used in production. On the other hand, SGA is *"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, delivery, marketing, research, and development, among others.

Table 1 presents summary statistics of the customer-firm matched data. Nielsen-GS1 data reveals that much of the firm-group-year sales arises from the number of customers, not from average sales per customer. More than 500,000 customers spend only approximately \$10 for each product group and firm per year on average. Among the total number of customers firms face in product group g, approximately half of them are newly purchasing the products every year. Also, the Nielsen-Compustat data present a large heterogeneity in firms' cost structure despite its smaller sample size relative to the Nielsen-GS1 data. Documenting the SGA-to-OPEX ratio shows that the non-production cost is approximately 30% of the total cost on average and is an important component of firms' total cost. Although we only have 332 firms in the Nielsen-Compustat matched data, they cover close to one-fourth of total sales in Nielsen. Appendix A provides a description of the cleaning procedure along with a more detailed description of the data.

2.2 Empirical Facts

2.2.1 Firm Sales Growth

We show that the acquisition of new customers is of first-order importance for firm sales growth. As firms enter the economy, there are two ways for them to increase their sales based on Equation (2.1); they can either raise sales for each customer or increase the total number of customers they face. To quantify the importance of each margin for firm sales

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
$p_{igt}D(p_{igt})$	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
$Capital_{it}$	1,658	2,541.30	8,045.80	2.17	232.01	5,947.71
SGA-to-OPEX$_{it}$	2,101	0.30	0.20	0.08	0.27	0.58
Sales-to-COGS$_{it}$	2,299	1.79	1.06	1.14	1.49	2.61

Table 1: Summary Statistics

Notes: The Nielsen-GS1 data in Panel A has 40,418 firms and 109 product groups in the period of 2005-2016. S_{igt} is sales of firm i in product group g and time t, $p_{igt}D(p_{igt})$ is sales per customer, m_{igt} is number of customers, m_{igt}^{New} is new customers in year t, and m_{igt}^{Old} is 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 in thousand 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. 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; other Compustat variables are deflated with the GDP deflator.

growth, we estimate the following equation:

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

where S_{it} is sales, and its components, of firm i in year t, age$_{it}$ is the number of years firm i stayed in the economy after the first 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). δ_a are the parameters of interest that measure the average sales and its components for each firm age group by using the variations within the firm and year. To capture firm entry, we only consider those firms that appear for the first time after a benchmark period between 2004 and 2007.[7]

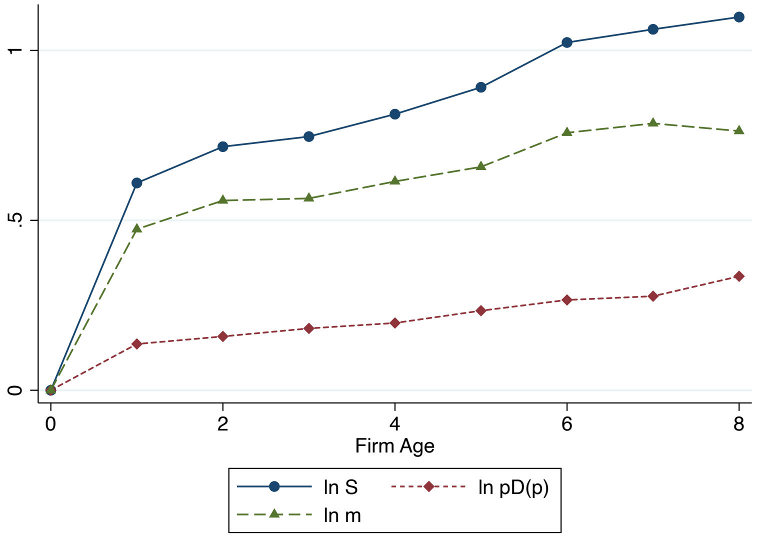


Figure 1: Decomposition of Firm Sales Growth by Firm Age

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), which include the firm and year fixed effects. The blue circle 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 is 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, firm sales growth is mostly attributed to the increase in the number of customers, not to the increase in sales per customer. Figure 1 plots the average log sales as a function of firm age (δ_a) and exactly decomposes it into the average log number of customers and the average log sales per customers. 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 get older, on average it still explains approximately 70% of sales for the maximum firm age observed in the data. The results are generally robust to using only those firms that appear at least 3 or 5 consecutive years or using average monthly sales, as shown in Appendix B.

In addition to decomposing firms' sales growth over time, we show that the underlying source of variation in firm-group-year sales in the cross-section of all firms in the Nielsen dataset is the variation in the number of customers (m_{igt}), rather than sales per customer

[7] 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 benchmark period to make the analysis conservative. The maximum firm age is eight since we use all years available in the Nielsen-GS1 data, 2004-2016.

($p_{igt}D(p_{igt})$). Following Equation (2.1), Table 2 decomposes the variance of log sales into log sales per customer, the log number of customers, and the covariance between these two components of sales. The number of customers accounts for approximately 80% of the variation of sales across firms, the sales per customer accounts for approximately 11% of the variance, and the covariance account for the rest.

$Var_i(lnS_{igt})$	$Var_i(lnp_{igt}D(p_{igt}))$	$Var_i(lnm_{igt})$	$2Cov_i(lnp_{igt}D(p_{igt}), lnm_{igt})$
7.5807	0.8672	6.1146	0.5989

Table 2: Decomposing the Variance of Sales

Notes: S_{igt} is sales, $p_{igt}D(p_{igt})$ is sales per customers, and m_{igt} is 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.

2.2.2 Non-production Costs

Given the importance of the number of customers in firm sales growth, a natural question arises: can firms affect the speed of growth of their customer base? One natural candidate that could explain differences in firms' number of customers is their non-production expenses, such as advertising and marketing costs. Here, we provide empirical evidence that shows that non-production costs are associated with the number of new customers firms face.

As a first step, we establish that the non-production 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. We empirically assess the plausibility of these assumptions by analyzing the correlation between sales and the non-production costs. In particular, to ease the concerns related to omitted variables, we include firm and year fixed effects and compare the co-movement of non-production costs and sales with the co-movement of other costs that in the literature are commonly assumed to be variable and fixed in the short-run: production costs and capital.

We estimate the following equation:

$$\Delta \ln(\text{Cost}_{it}) = \beta \Delta \ln \text{S}_{it} + \lambda_i + \lambda_t + \varepsilon_{it}, \tag{2.2}$$

where Cost$_{it}$ denotes SGA expenditure (non-production costs), COGS (production costs), or capital expenditure of firm i in quarter t. λ_i and λ_t are firm and quarter fixed effects, respectively. The coefficient β measures the co-movement of different types of costs with sales. Since almost all costs become variable costs in the long-run, we use quarterly data from Compustat, which is the shortest time-frequency available. Thus, our analysis can be interpreted as providing a lower bound for the degree of variability of different costs at longer horizons. In addition, we measure the co-movement with small changes in sales across two consecutive quarters to ease concerns about one-time changes in fixed costs that could be correlated with sales.

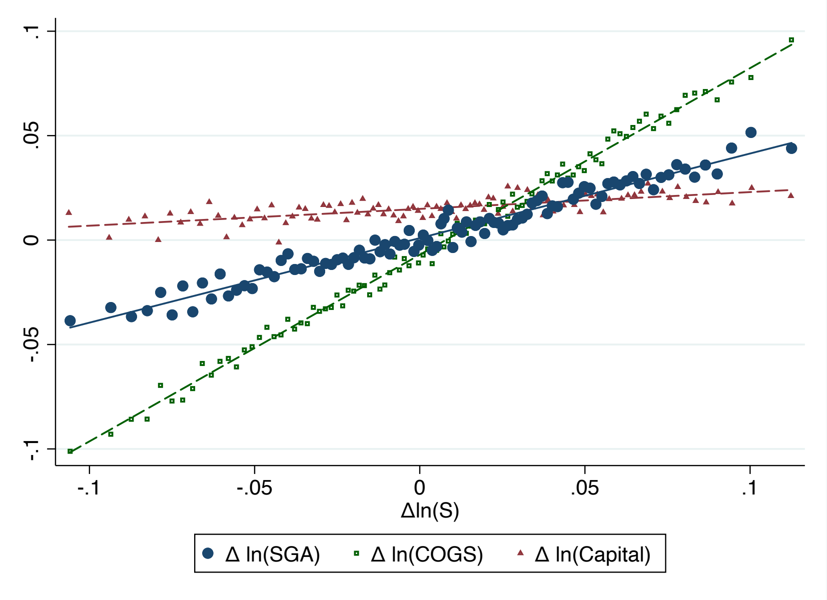


Figure 2: The Semi-variable Nature of SGA

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). There are 17,168 firms in 1964-2016 used in this analysis.

We find that non-production costs exhibit significant co-movement with sales in the short-run. According to this metric, non-production costs appear to be more variable relative to changes in the capital stock and less variable relative to production costs. Figure 2 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.[8] 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

[8] Table C.2 in the Online Appendix presents the regression results and Figure C.4 plots the results for the entire range of $\Delta \ln S_{it}$. All of them show very similar patterns on the variability of different types of costs.

($\beta = 0.894$; SE 0.008), and the lowest co-movement with investment ($\beta = 0.081$; SE 0.008).[9] Non-production costs exhibit intermediate levels of co-movement with sales ($\beta = 0.450$; SE 0.010).

Throughout our analysis, we focus on the variable component of SGA that is associated with firms' sales in the short-run. The semi-variable nature of SGA suggests that non-production costs consist of both a variable and a fixed component. The variable costs include advertising and marketing expenses that can increase firms' annual customer bases. The fixed costs could contain other components, such as Research and Development (R&D) expenses, that affect firm behavior in the long-run. In fact, similar to capital expenditures, R&D costs exhibit weak correlation with short-run sales, as shown in Appendix B.

Given the positive correlation between a firm's non-production costs and total sales, we next exactly decompose this correlation into two different components by leveraging the Nielsen-Compustat matched data: the correlation of non-production costs with (i) the number of customers and (ii) sales per customers. For this, we estimate the following specification:

$$\ln \mathrm{S}_{igt} = \gamma_1 \ln \mathrm{SGA}_{it} + \boldsymbol{X'_{it}\gamma_2} + \lambda_{ig} + \lambda_{st} + \lambda_{gt} + \varepsilon_{igt},$$

where S_{igt} is sales and its components of product group g that firm i sells in year t, $\boldsymbol{X'_{it}}$ is a vector of firm-time-level control variables, λ_{ig} are firm-group fixed effects, λ_{st} are 2-digit SIC-year fixed effects, and λ_{gt} are 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 $\boldsymbol{X'_{it}}$ 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_1, which exactly decomposes the correlation between total sales and SGA into the correlations of each component of sales and SGA.

As shown in Table 3, almost all of the correlation of sales and non-production costs arises from the correlation of the non-production costs with the number of *new* customers, not with the sales per customer or the number of existing customers. Column (1) confirms the results shown in Figure 2 in a different specification and sample, and columns (2) and (3) decompose the results in column (1). Approximately 95% (0.090/0.095) of the correlation between sales and SGA is attributed to the correlation between the number of customers and SGA. Columns (4) and (5) further decompose the number of customers into the number of new customers and old customers. While there is a strong correlation between SGA expenses

[9] The empirical estimates are close to what is in simple theoretical models. For example, in the simplest model with competitive markets and Cobb-Douglas production function, the slope of equation Equation (2.2) for (production) labor costs should be equal to one. If capital is assumed to be fixed in the short-run, then the slope for the user cost of capital is mechanically equal to zero.

	Decomposition of ln S			ln m: New vs. Old	
	(1)	(2)	(3)	(4)	(5)
	ln S	ln pD(p)	ln m	ln $\mathrm{m}^{\mathrm{New}}$	ln $\mathrm{m}^{\mathrm{Old}}$
ln $\mathrm{SGA}_{\mathrm{it}}$	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 3: 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 is sales, $pD(p)$ is sales per customers, m is number of customers, m^{New} is number of new customers, and m^{Old} is 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 ($\mathrm{m} = \mathrm{m}^{\mathrm{New}} + \mathrm{m}^{\mathrm{Old}}$). SIC is a two-digit SIC code, and all Nielsen variables are projection-factor adjusted.

and the number of new customers, the regression coefficient for old customers is neither economically nor statistically significant. These results show that the non-production costs of firms are associated with the acquisition of new customers, rather than maintaining the existing customer base.

2.2.3 Price-Cost Markups

Armed with the empirical evidence showing that big firms in the economy mostly arise from the endogenous acquisition of new customers, 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). Canonical models with endogenous markups predict that markups are a function of the market share of firms, implying that large firms charge higher markups. We revisit the link between the firm-level markups and market share by estimating the following regression equation:

$$\ln \mathrm{Markup}_{it} = \alpha_1 \ln \mathrm{pD(p)}_{it} + \alpha_2 \ln \mathrm{m}_{it} + \lambda_t + \lambda_s + \varepsilon_{it},$$

where Markup_{it} is price-cost markup of firm i at time t, which is measured as Sales-to-COGS ratio following previous studies (see, e.g., Traina (2019) and De Loecker et al. (2020)).[10] The

[10] Price-cost markups can be recovered as the output elasticity with respect to COGS divided by the COGS share of Sales from the firm's first-order condition with respect to COGS (see e.g., De Loecker et al., 2020).

sector-year fixed effects $\lambda_{s,t}$ absorb all the variation at the sector-year-level, which allow us to interpret a firm's sales as its market share.

	ln Markup				
	(1)	(2)	(3)	(4)	(5)
ln pD(p)_{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 4: 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. The markups are measured as sales-to-COGS ratio. SIC is a two-digit SIC code, and all Nielsen variables are projection-factor adjusted.

We find that firms' markups are highly correlated with their sales per customer but are unrelated to their number of customers, as reported in Table 4. Results are robust to including different combinations of year and sector fixed effects. Appendix B presents the regression results by using firms' sales instead of the number of customers, and the greater importance of sales per customer in explaining markups remains. Given the inclusion of industry-time fixed effects, we can interpret our results as showing that firms that charge higher markups have a higher market share in terms of sales per customer, and not a higher market share in terms of the number of customers. That is, the relevant notion of market share is based on sales per customer.

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.

If the output elasticity is constant across industries and time, then the Sales-to-COGS ratio identifies all the variation in markups. If instead the output elasticity is assumed to be industry-, or time-, or industry-time-specific, then the industry-, time-, or industry-time fixed effects included in the regression absorb all the variation in the output elasticity, and the Sales-to-COGS ratio identifies the variation in markups.

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.

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. In particular, each member of the household is matched to, and forms demand for, one and only one variety at each given time. We let $m_{i,t}$ denote both the measure and the set of matches for variety i, and write $j \in m_{i,t}$ when member j is matched to variety i.

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

$$\max_{\{C_t, L_t, (c_{i,j,t})_{j \in m_{i,t}}^{i \in N_t}\}} \sum_{t=0}^{\infty} \beta^t \left[\frac{C_t^{1-\gamma}}{1-\gamma} - \xi \frac{L_t^{1+\psi}}{1+\psi} \right] \tag{3.1}$$

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

$$\int_0^{N_t} \int_0^1 \mathbf{1}_{\{j \in m_{i,t}\}} p_{i,t} c_{i,j,t} djdi \leq W_t L_t + \int_0^{N_t} \Pi_{i,t} di - T_t$$

Here, $c_{i,j,t}$ is the consumption of household 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 supplied by 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 members that are matched to a particular variety i choose to have the same level of consumption from that variety, whereas the members that are not matched to i consume zero:

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

Here,

$$D_t \equiv \left[\int_{i \in N_t} \int_0^1 \mathbf{1}_{\{j \in m_{i,t}\}} \frac{c_{i,j,t}}{C_t} \Upsilon'\left(\frac{c_{i,j,t}}{C_t}\right) djdi \right]^{-1}$$

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

is an *aggregate demand index* and P_t is 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 matches 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 match* of variety i. The expression for total demand 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 the number of customers times the sales of firms per customer. In our model, the same decomposition holds:

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

Elasticities and Super-elasticities of Demand Following Edmond et al. (2018), we use the 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] Implementing this functional form in Equation (3.3), we can derive the following expression for the relative demand per match 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

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

$\sigma > 1$ and $\eta \geq 0$ that vary the *elasticity* and *super-elasticity* of demand. Formally, these quantities are given by

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

Here, $\varepsilon_{i,t}$ is the expression for the *elasticity of demand* and shows that under the Klenow and Willis (2016) specification, demand elasticity is a *decreasing* function of the relative demand per match. 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. 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.[14]

Dynamics of Matches We assume that every household member at each given period is matched to one and only one operating firm.[15] Two processes in the model cause separation: (1) at the end of each period, customers of exiting firms separate and become available for new matches, and (2) customers of incumbent firms separate at an exogenous rate of $\delta \in [0, 1]$.

Firms launch marketing campaigns to attract new customers from the pool of newly separated consumers. In particular, 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 matches for firm i at t is proportional to the number of ads that they posted, but it is normalized by the total number of ads and 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 matches should the firm choose to exercise that option. We interpret $P_{m,t}$ as the *cost of a match*. It is the number of ads that a firm needs to post to get one new customer and is determined so that the

[14] It is also straightforward to show that when $\eta \to 0$ the relative demand function in Equation (3.4) approaches to the familiar CES demand function. To see this note that with $\eta \to 0$, $D_t(1 - \sigma^{-1}) = 1$ and

$$\ln(q_{i,t}) = \lim_{\eta \to 0} \frac{\sigma}{\eta} \ln(1 - \eta \ln(p_{i,t})) = -\sigma \ln(p_{i,t}).$$

[15] This implies that while advertising changes the distribution of customers across firms in an industry, it does not increase the total number of household members that buy from that industry.

matching market clears:

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

This expression shows that the cost of a match decreases with the total number of separated members 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 3. 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 marketing 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:

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

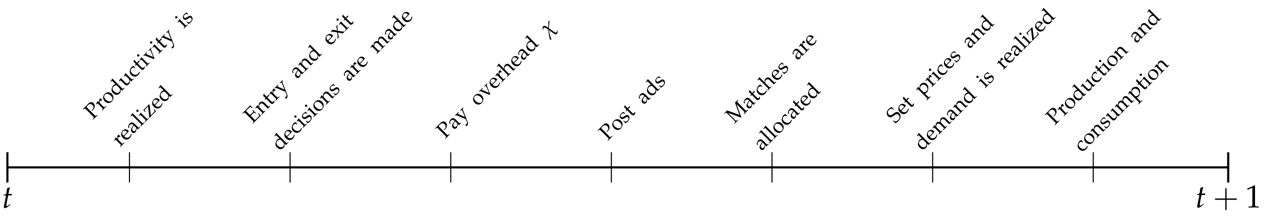
We let Λ_t denote the set of these potential entrants at t. Incumbents, firms that entered the economy at least one period ago, also draw new productivities according to the following AR(1) process: [16]

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

With new productivities drawn, each incumbent or potential entrant then decides whether or not to stay in the economy or dropout. We refer to this decision by $\mathbf{1}_{i,t} \in \{0,1\}$, with 1 being an indicator for entering or staying. Finally, of all the firms who decided to stay, each draws a Bernoulli survival shock, $\nu_{i,t}$, that is 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.

[16] It is important to note that, 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).

Figure 3: Timing of Events


Productivity is realized
Entry and exit decisions are made
Pay overhead χ
Post ads
Matches are allocated
Set prices and demand is realized
Production and consumption
t
$t+1$

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

Marketing, 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 marketing. The firm's customer base then evolves according to the law of motion for matches per Equation (3.5), where $m_{i,t-1} \equiv 0$ for operating firms that entered at time t.

Furthermore, for a given number of matches, a firm's demand is given by Equation (3.2). Firms take this demand schedule as given and choose the price that maximizes their profit. They then produce to meet their realized demand using technology

$$y_{i,t} = z_{i,t} l_{i,p,t}^{\alpha},$$

where $z_{i,t}$ is the firms' productivity, and $l_{i,p,t}$ is the labor demand of the firm for production.

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

$$\begin{aligned} & v_t(m_{i,t-1}, z_{i,t}) \qquad (3.8) \\ \equiv & \max_{(p_{i,\tau}, l_{i,s,\tau}, l_{i,p,\tau}, \mathbf{1}_{i,\tau})_{\tau=t}^{\infty}} \mathbb{E}_t \sum_{\tau=t}^{\infty} (\beta\nu)^{\tau-t} \left(\prod_{h=t}^{\tau} \mathbf{1}_{i,h} \right) \left(\frac{C_\tau}{C_t} \right)^{-\gamma} \left[\underbrace{p_{i,\tau} y_{i,\tau}}_{\text{Total sales}} - \underbrace{W_\tau l_{i,p,\tau}}_{\text{COGS}} - \underbrace{W_\tau (l_{i,s,\tau} + \chi)}_{\text{SGA costs}} \right] \end{aligned}$$

subject to

$$y_{i,\tau} = m_{i,\tau} q_{i,\tau} C_\tau = z_{i,\tau} l^\alpha_{i,p,\tau} \tag{3.9}$$

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

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

That is, 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 by producing enough, Equation (3.10) that specifies demand per match 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 matches and the non-negativity of labor allocated towards marketing.

3.3 Characterization of Firms' Decisions

In this section, we characterize the firms' optimal decisions rules for pricing, marketing and entry/exit.

Prices and Markups For a firm that has decided to operate in a given period, and for a given choice of marketing that determines its number of matches 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 the common proportionality relationship derived between the labor share and markup in conventional models also holds in our model, which verifies our measurement of markups in the data.

It is also important to note that $\varepsilon_{i,t}$, the firm's elasticity of demand, itself is a function of demand per match 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, in regards to which we can derive the following result.

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 C.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 marginal cost, the 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—this is in particular in contrast to 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 4.

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 C.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 marketing 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 C.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 as long as variable elasticities are assumed. However, in our model, firms can be big either because they sell more per customer—hence, charging higher markups—or because of having a larger number of customers, which translates into lower markups once we control for productivity and sales per customer.

Marketing Strategies A key feature of our model is that firms internalize the decision of generating new matches and can spend resources to do so. For this decision, while the marginal cost of generating a match 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 match. The following proposition formulates the optimality condition for firms' marketing decision in terms of this cost-benefit analysis.

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

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

Proof. See Appendix C.4. □

Equation (3.16) shows that the marginal benefit of a 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 $\delta = 1$, in which case the customer base $m_{i,t}$ is given by

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

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 higher productivity than $z^*(m_{-1})$ stay or enter the economy.

3.4 Equilibrium

We are now ready to define an equilibrium concept for this economy under monopolistic competition.

Definition 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:

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

4. the set of operating firms, N_t, follows

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

Solution Method We solve the model globally by combining collocation methods and nonstochastic simulation (see Young, 2010) to approximate the distribution of firms. Appendix D 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

Given an initial distribution of productivity in the economy, 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 base if the firm has too many customers or by launching marketing campaigns if the firm needs to grow—and (3) and how much each operating firm should produce. A formal statement of the Planner's problem is included in Appendix C.5.

Given the planner's problem, our objective here is to characterize the efficient allocation, especially for the distribution of customers across firms,which is unique to our setting relative to other models of variable markups. A second source of inefficiency that emerges in our model is the war-of-attrition nature of advertisement that leads to overuse of labor for marketing in the equilibrium. In order to focus on the misallocation of customers, we will abstract away from this second source of inefficiency by assuming that the social planner has to spend the same amount of aggregate labor for matches as in the equilibrium. The following Lemma shows that this assumption is without loss of generality for the optimal distribution of customers.

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

Proof. See Appendix C.6. □

This result follows from the fact that returns to marketing are fully relative in labor allocated towards posting ads. Moreover, since any distribution of matches can be implemented by any given amount of aggregate labor, Lemma 2 also implies that all distributions of matches have the same cost for the planner. Therefore, 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 members based on firms' new productivities.

To characterize the efficient allocation, there are two margins that need to be considered. First, fixing the 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 C.7. □

Proposition 3 shows how our model breaks the link between size and misallocation. The planner would like all *matched* members to have the same level of consumption from the varieties that they are matched to and capitalize on the higher efficiency of more productive firms by giving them more matches. In other words, it is efficient for all operating firms to have *equalized sales per match*, with more productive firms having more matches. 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 matches than what is efficient).

Our result in Proposition 3 is also at odds with the trade-off that the social planner faces in the 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*. On the other hand, dispersed *relative consumption* of different varieties is inefficient due to the weak substitutability of goods. In those models, since $m_{i,t}$ is exogenous, the social planner has to balance these two forces only by choosing the distribution of $q_{i,t}$. In our model, the social planner equalizes relative consumption across all matched members and increases the aggregate consumption of the economy by giving more customers to more productive firms.

Recognizing the endogeneity of the allocation of matches has two important macroeconomic implications. First, the Pareto frontier of the economy is broader than what is implied by conventional models—since the social planner in our model always has the option to replicate the homogeneous allocation of matches. Second, the magnitude of losses from misallocation, and the distance of the equilibrium allocation from this new frontier, depends on how effective the equilibrium matching 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 model.

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

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

$$G_t^*(z_{i,t}) \equiv \max_{\{(\mathbf{1}^*_{i,\tau})_{\tau\geq t}\}} \mathbb{E}_t \sum_{\tau=t}^{\infty} (\beta\nu)^{\tau-t} \left(\prod_{h=t}^{\tau} \mathbf{1}^*_{i,h}\right) \left(\frac{C^*_\tau}{C^*_t}\right)^{-\gamma} \left[\underbrace{m^*_{i,\tau}C^*_\tau}_{\text{Total sales}} - \underbrace{m^*_{i,\tau}W^*_\tau L^*_{p,\tau}}_{\text{COGS}} - \underbrace{W^*_\tau \chi}_{\text{SGA costs}}\right],$$

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 matches given to a firm with productivity $z^*_{i,t}$.

Proof. See Appendix C.8. □

It then follows that the planner only keeps a firm if this value is strictly positive. An important observation here is that the social value of a firm is strictly decreasing in $m^*_{i,t}$ which is much more sensitive to $z_{i,t}$ than in the equilibrium. In our model, the planner wants a significant amount of concentration of matches at the top of the productivity distribution. Relative to the conventional model, in our model, firms with lower productivity have lower social value because it is optimal for them to serve fewer customers under the efficient allocation.

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 an *aggregated production function* as well as an *aggregate markup* that characterizes the wedge between the aggregate marginal product of labor and the wage—which is equal to the marginal rate of substitution once we impose the optimal labor supply condition of the household. 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 of this economy, 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 uses Equation (3.9) that requires the firm to hire enough labor to produce and meet its total demand.

Defining the aggregate output of this economy as $Y_t \equiv C_t$—since all aggregate output is consumed by the household—and rearranging Equation (3.17), we arrive at the aggregate production function of this economy 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 its marginal cost—the wage W_t. Formally, having derived the aggregate production function in Equation (3.18), the aggregate markup is defined as

$$\mathcal{M}_t \equiv \frac{\partial Y_t / \partial L_{p,t}}{W_t} = \alpha^{-1} \frac{Y_t}{W_t L_{p,t}} \tag{3.20}$$

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

$$\mu_{i,t} \equiv \alpha^{-1} \frac{p_{i,t} y_{i,t}}{W_t l_{i,p,t}},$$

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 cross-multiplying this equation, integrating over i and dividing it by Equation (3.20), we can then derive that the aggregate markup is the production cost-weighted average of firm level markups—an expression akin to the one derived in Edmond et al. (2018):

$$\mathcal{M}_t = \int_{i \in N_t} \omega_{i,t} \mu_{i,t} di, \tag{3.21}$$

where

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

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.

Decompositions The following proposition shows that for small perturbations around the equilibrium allocation, we can write the welfare losses of the households as a function of changes in the aggregate TFP in Equation (3.19) and changes in labor supply.

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*

$$\underbrace{\frac{\Delta U_t}{U_{c,t}C_t}}_{\Delta \textit{Welfare (C.E.)}} \approx \underbrace{\Delta \ln(Z_t)}_{\Delta \textit{TFP}} + \underbrace{\alpha(1-\mathcal{M}_t^{-1})\Delta \ln(L_{p,t})}_{\Delta \textit{Losses from Aggregate Markup}} - \alpha \mathcal{M}_t^{-1}\left[\underbrace{\chi \frac{N_t}{L_{p,t}}\Delta \ln(N_t)}_{\Delta \textit{Losses from Entry/Exit}} + \underbrace{\frac{L_{s,t}}{L_{p,t}}\Delta \ln(L_{s,t})}_{\Delta \textit{Losses from Advertising}}\right] \tag{3.22}$$

where Z_t is the aggregate TFP in Equation (3.19), $\mathcal{M}_t$ is the equilibrium aggregate (cost-weighted) markup in Equation (3.21), $L_{p,t}$ 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 C.9. □

Equation (3.22) 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 labor supply that are allocated towards the overhead costs of operating firms, and (4) changes in labor supply that are 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, holding the fourth channel fixed.[17]

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 by matching, via the Simulated Method of Moments, several micro- and macro-moments related to firm dynamics in the US economy in 2012.

[17] As we discussed in deriving the efficient allocation, since advertising is relative and hence a pure war-of-attrition in our model, all labor that is allocated towards advertising is inefficient from a social perspective. While fixing the amount of labor allocated towards advertising across different allocations is a choice on our part and could be easily relaxed, it constitutes a clean benchmark that allows us to focus on quantifying the main channel of interest in our analysis, the misallocation of demand across firms.

4.1 Calibration Strategy

The most relevant parameters to calibrate are the size of fixed overhead costs χ and the elasticity of the matching function ϕ. Both parameters affect not only the overall size of SGA relative to total costs, but also the composition within SGA between its variable and fixed components.[18] As we show in Appendix B.1, the average COGS-to-OPEX ratio exhibits a significant size profile, based on firms' sales. Therefore, it is important to also have a good fit of the sales distribution in the economy. Ideally, we would combine aggregate data on the size distribution of firms with aggregate data on firms' cost structure. The fact that the latter is only available for a subset of firms from Compustat poses a challenge, which we attempt to solve 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.

Fixed Parameters We set the model 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 $\alpha = 0.64$. During the calibration exercise, we normalize the measure 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.[19] 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 using Nielsen

[18] Note that our model does not impose the existence of a variable SGA component, nor the fact that firms *need* to spend resources to acquire customers. 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 case of pure fixed SGA costs and exogenous customer acquisition as a special case.

[19] 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.

data an annual retention probability of 0.85 for the top two largest firms in each product category. If we also restricted 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.[20] 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 the set of moments chosen, below we provide a discussion of which moment should intuitively be more relevant to identify each parameter. We formalize this discussion in Appendix E 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 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 C.10. □

Equation (4.1) is obtained from the firm's optimality condition regarding customer ac-

[20] 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 D.2 provides the computational details of the calibration exercise.

quisition. 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 general case, Appendix E 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 both moments 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).[21]

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 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 proportional to the production markup $\mu_{i,t}$. The following proposition shows that a similar relationship holds in a special case of the model.[22] Appendix E 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 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 C.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

[21] 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.

[22] 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).

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.[23]

The parameter of the AR(1) productivity process for incumbent firms, ρ_z and σ_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 (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 elasticity of 0.474. This similarity lends support to the identification argument provided in Proposition 6.

Table 6 and Figure 4 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 4 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 E, we show the data and model relationships between 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

[23] 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. 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 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 chosen externally. Panel B contains parameters internally calibrated to match moments presented in Table 6 and Figure 4.

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, sales per customer and the size of the customer base.

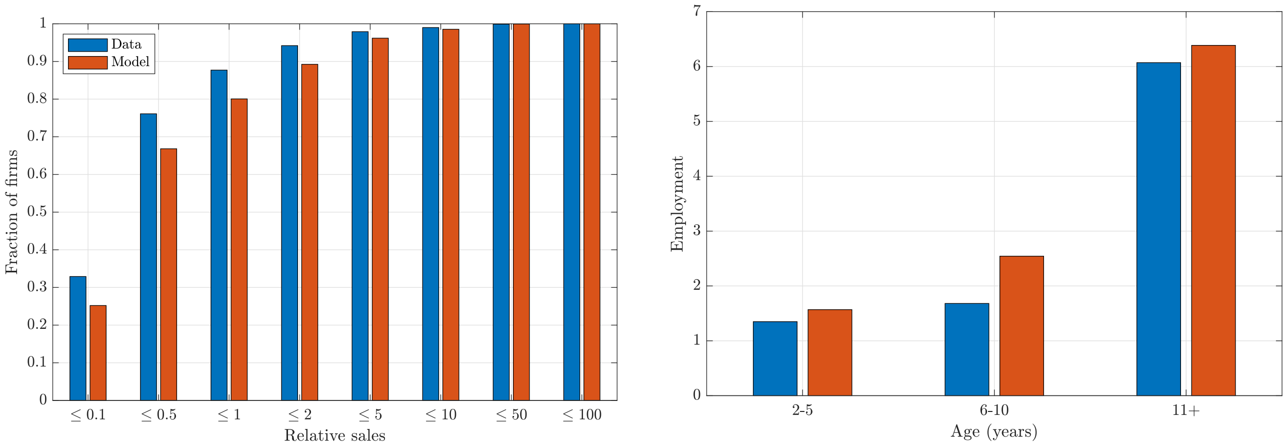
Firm Dynamics Figure 5 shows two model moments that were not explicitly targeted during the calibration exercise: the average exit rate by age and the average employment growth by age. As Panel A shows, the average exit rate is decreasing in the firm's age, as in the data (see e.g., Haltiwanger, Jarmin and Miranda, 2013). The fact that entrants enter the economy with lower average productivity and no customer base makes young firms

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 $SGA_{i,t} = \beta Sales_{i,t} + \psi COGS_{i,t} + \varepsilon_{i,t}$. Avg. COGS-to-OPEX ratio refers to cross-firm average of the ratio. 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 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 4: Model Fit: Firm Size



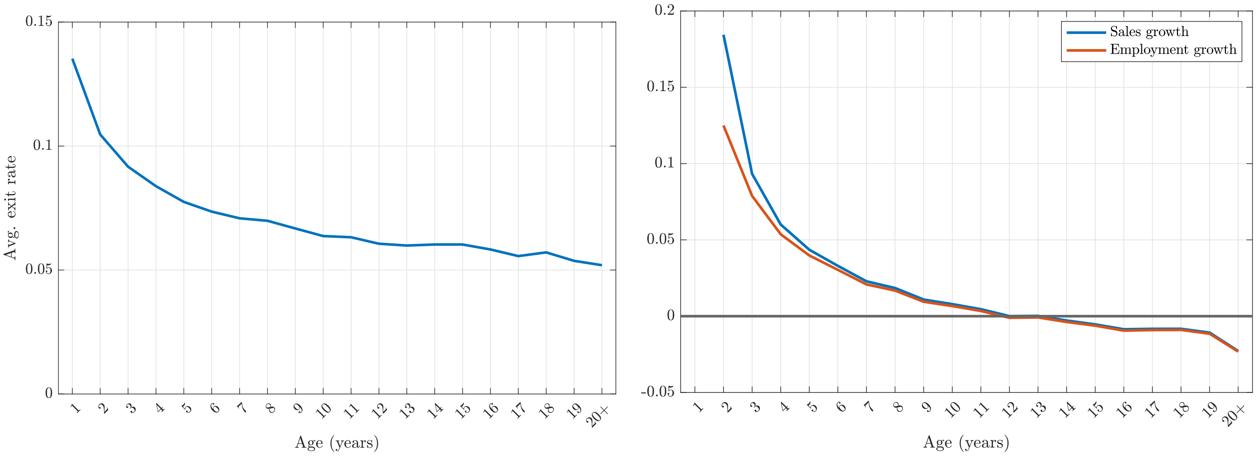
(a) Distribution of Sales

(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.

more likely to exit 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 calibrate the size of the overhead cost χ. Here, we took another route by 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 a decreasing average employment growth as a function of firms' age. Both patterns are consistent with the empirical evidence in Haltiwanger et al. (2013). For example, the average net employment growth rate of 1-2 years old and 7-8 years 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-years old firms are 12.5% and 2.1%, respectively.

Figure 5: Exit and Growth by Age



(a) Exit Rate by Age

(b) Avg. Growth by Age

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 explain 11.4% of the variance of 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 only driven by productivity shocks, the model over-predicts the size of the covariance term.

Table 7: Sources of Dispersion in Sales 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 report the variance of the log sales per customer, $var\left(\ln p_{i,t}\Upsilon'(p_{i,t}/D_t)\right)$. The second column report the variance of the log number of customers, $var\left(\ln m_{i,t}\right)$. The last column report the covariance between both terms, $cov\left(\ln m_{i,t}, \ln p_{i,t}\Upsilon'(p_{i,t}/D_t)\right)$. 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 the main source of market power (see Table 2.2.3). 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

$$\ln(\mu_{i,t}) = \theta_0 \ln\left(p_{i,t}\Upsilon'\left(\frac{p_{i,t}}{D_t}\right)\right) + \theta_1 \ln m_{i,t} + \kappa_t + \varepsilon_{i,t}.$$

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 a 1% increase in the average sales per customer is associated with a 0.11% increase in markups. This point estimate barely lies outside the 95% confidence interval estimated in the data. On the other hand, a 1% increase in the customer base, increases markups by only 0.02%. The only reason why this coefficient is not 0 in the simulated data is the nonlinear nature of the relationship between these variables. If we instead conditioned on a flexible function of sales per customer, then a firm's number of customers should have no predictive power for markups, as in the data.

5 The Role of Endogenous Customer Acquisition

The goal of this Section is to investigate the role played by endogenous customer acquisition, both in directly shaping a firm's optimal choices and indirectly determining the aggregate properties of the equilibrium. To do so, we compare the allocations and equilibrium of the calibrated model (labeled as "Baseline" from here on) with an alternative model in which each period firms receive a fixed number of total customers without having to spend

Table 8: Sources of Dispersion in Sales and Markups

	ln $\text{Markup}_{\text{i,t}}$	
	Data	Model
ln sales per $\text{customer}_{\text{i,t}}$	0.059***	0.111
	(0.022)	
ln n. of $\text{customers}_{\text{i,t}}$	0.002	0.022
	(0.007)	
Observations	2433	
R^2	0.313	0.869
Year FE	✓	✓
SIC FE	✓	

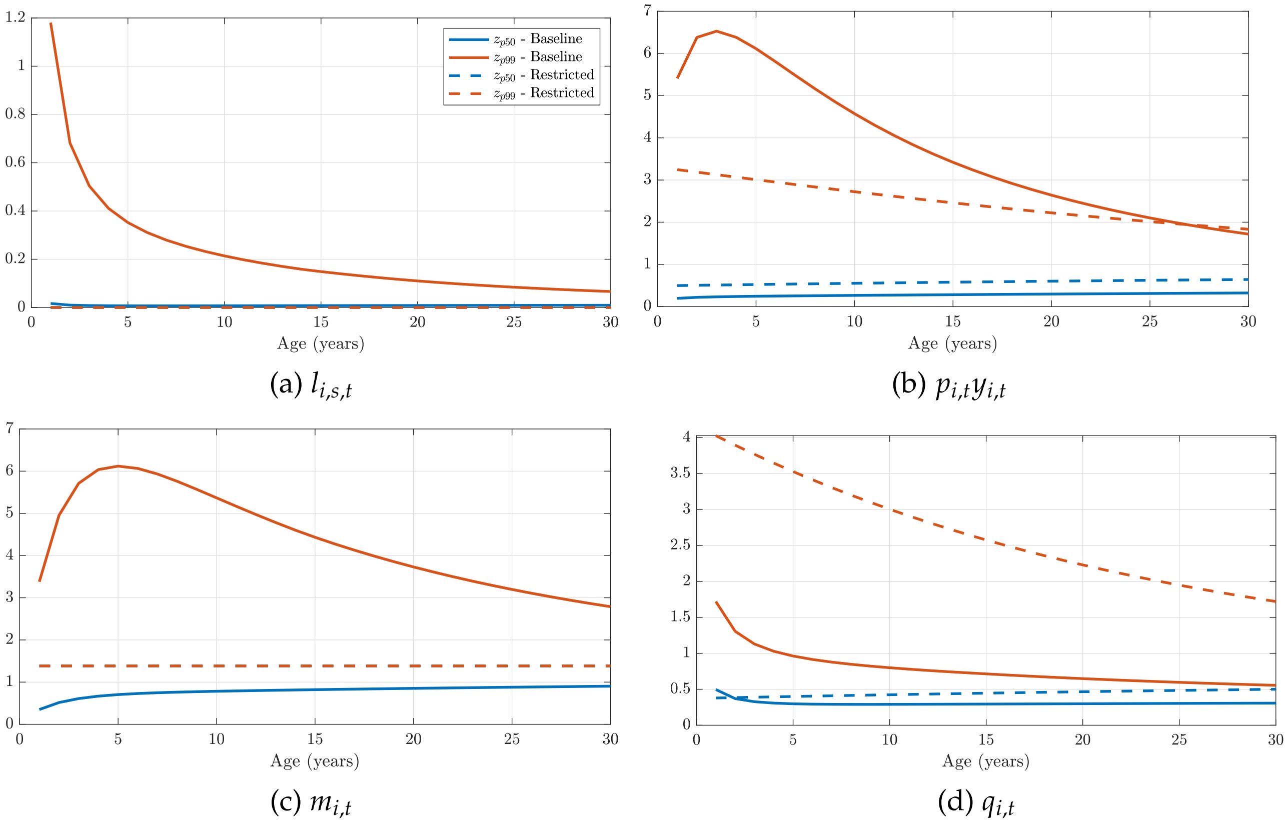
Notes: This table reports the results of an OLS regression of a firm's log markup ($\ln(\mu_{i,t})$) on log sales per customer ($\ln\left(p_{i,t}\Upsilon'\left(p_{i,t}/D_t\right)\right)$) and log size of the customer base ($\ln m_{i,t}$). Column (1) reproduces the empirical estimates from Table 4. Column (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.

any resources (labeled as "Restricted" from here on).

Implications for Concentration Figure 6 plots the firm dynamics for two firms that start with zero customer base and 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 firms front load their efforts to acquire customers when young and that spending in $l_{i,s,t}$ is higher for more productive firms. Panel B shows that the sales of the more productive firm are larger than the ones of the less productive firm, increase over time due to a larger stock of customers and finally decrease due to mean-reverting productivity. How do firms increase their sales? Panels C and D 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 spending in customer acquisition. On the other hand, sales per customer are higher when firms are younger, due to mean reverting productivity and decreasing returns 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 model with exogenous customer base, 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 model with exogenous customer base. Thus, relative to that model, endogenous customer acquisition increases the dispersion of customer bases 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 concentrate 17% of sales, in our baseline model they concentrate 50% of sales.

Figure 6: Average Firm Dynamics



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

(b) $p_{i,t}y_{i,t}$

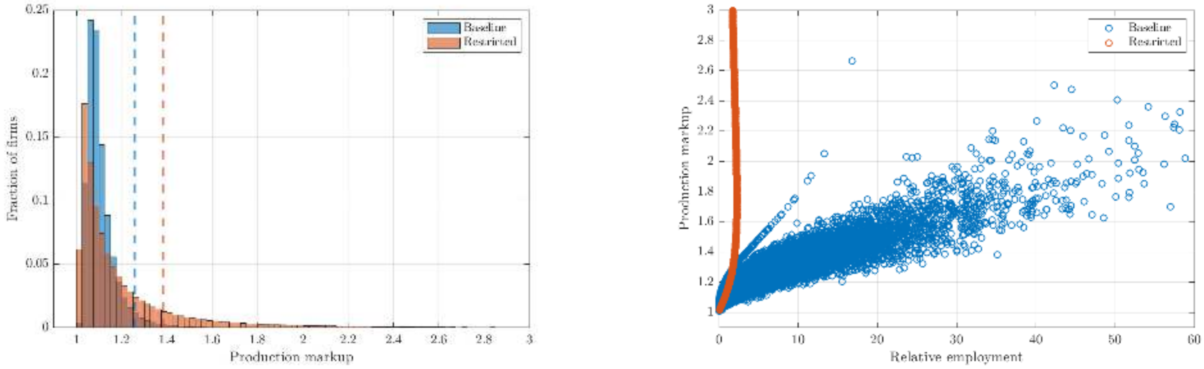
(c) $m_{i,t}$

(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 7 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 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.

Figure 7: Customer Acquisition and Market Power



(a) Distribution of Markups

(b) Avg. Size by Age

Notes: Panel (a) plots the distribution of production markups in the baseline and restricted models. 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 ($m_{i,t} = 1/N_t$).

The following decomposition of the difference in aggregate markups, which follows from Equation (3.21), 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.[24] The first term (denoted "Δ Distribution") captures the contribution of differences in the distribution of relative employment across firms and keeps 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 and keeps the distribution of relative employment fixed at the 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

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

markups in the baseline model *reduces* the aggregate markup by 18.7pp, so the aggregate markup ends up being smaller by 9.71pp. Thus, why does the baseline model feature much higher concentration but a lower aggregate markup? This is 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.

Aggregate Implications Endogenous customer acquisition, by allowing consumers to be concentrated among high productivity firms, affects aggregate outcomes beyond market power. First, it allocates more of the economy's resources to more productive firms, which reflects itself in higher aggregate TFP. The way to measure this is through the cost-weighted TFP across firms (see Baqaee and Farhi, 2019; Edmond et al., 2018). 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 concentrate a higher share of production costs.

The distinction in our model between sales per customer and the number of customers also matters for aggregate output. In our model, more productive firms are big because of a larger customer base, which means that they are not restricted by their choke quantities that restrict consumption per customer and produce for a larger share of the population. We find that by imposing an exogenous customer base reduces output by 24 percent. Another implication is related to the number of operating firms in equilibrium. The flip side of the fact that in our model large firms have a larger customer base is that mechanically smaller firms sell to fewer customer, which reduces their profits 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.

Finally, Table 9 shows that in the restricted model aggregate 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 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.

6 Quantifying the Efficient Allocation

The objective of this section is twofold. First, we present and compare the difference between the equilibrium and efficient allocations and quantify the gains under the efficient

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.

allocation of customers and resources. In doing so, we also revisit our ex-ante decomposition of welfare in Proposition 5 and quantify the contribution of the three channels (TFP, Aggregate Markups and Overhead cost of Entry/Exit). 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:

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

There are two main takeaways from this decomposition: (1) the consumption equivalent welfare gains of the household under the efficient allocation is 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 labor allocated towards production, output is 14.6% higher under the efficient allocation but the number of firms is 11.3% lower and the 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.

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.

Dissecting the Gains in TFP To dissect 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*. The only subtle difference here is that in our comparison technological change only happens due to the compositional changes in the productivity distribution across the two allocations. Therefore, for us, technological change is a manifestation of the different entry and exit policy that the planner adopts for the economy. 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 terms 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 reallocation. 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.

It is important to note that the notion of shifting demand towards more productive firms is not new to our setting. Even in a model without customer acquisition, the planner would prefer to do this to some extent. However, in canonical models the planner faces a stronger trade-off: on one hand, shifting demand towards more productive firms increases allocative efficiency and allows for a higher level of aggregate consumption. On the other hand, since in canonical models the planner has to shift demand through relative consumption, $q_{i,t}$, higher concentration of demand across more productive firms comes with the inevitable cost of distorting the distribution of relative consumption due to the weak substitutability of goods. As a result, efficiency gains from reallocation are not large because the cost of distorting relative consumption is too high.

However, this trade-off for the planner is non-existent in our model because here the planner can equalize—and in fact does equalize—relative consumption across all varieties ($q^*_{i,t} = 1$), but allocates a larger number of customers towards more productive firms ($m^*_{i,t} \propto z_{i,t}^{\frac{1}{1-\alpha}}$). Another consequence of reallocating customers towards more productive firms is a 39.2% increase in concentration of sales among the top 5% of firms. In our model, since more customers only shift firms' demand without increasing their market power, higher concentration of customers across more productive firms is efficient to the point that marginal costs of production are equalized across all firms.[25]

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. 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, but it comes at the cost of having fewer 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 To study the forces at work here, we start by reviewing the usual costs and benefits of having more firms in the economy and then add the new

[25] In fact, as $\alpha \to 1$, the concentration of customers across firms becomes larger and in the limit all customers would be allocated towards the most productive firm.

mechanism that comes to 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 of the economy by allowing for higher *aggregate* returns to scale. 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.[26] 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 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. While in conventional models, in order to increase aggregate returns to scale, 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, the planner can achieve the same objective by allocating customers towards firms at the top of the productivity distribution. Moreover, since love for variety only matters through demand per customer (q), the higher concentration of customers at the top of the productivity distribution does not lead to a lower aggregate productivity through this channel. Hence, with this additional tool, our planner is able to achieve a higher returns to scale without having to pay for the overhead costs of more firms, which leads to a lower number of firms in the equilibrium and increases the welfare of the household by 1.6% as shown in Equation (6).

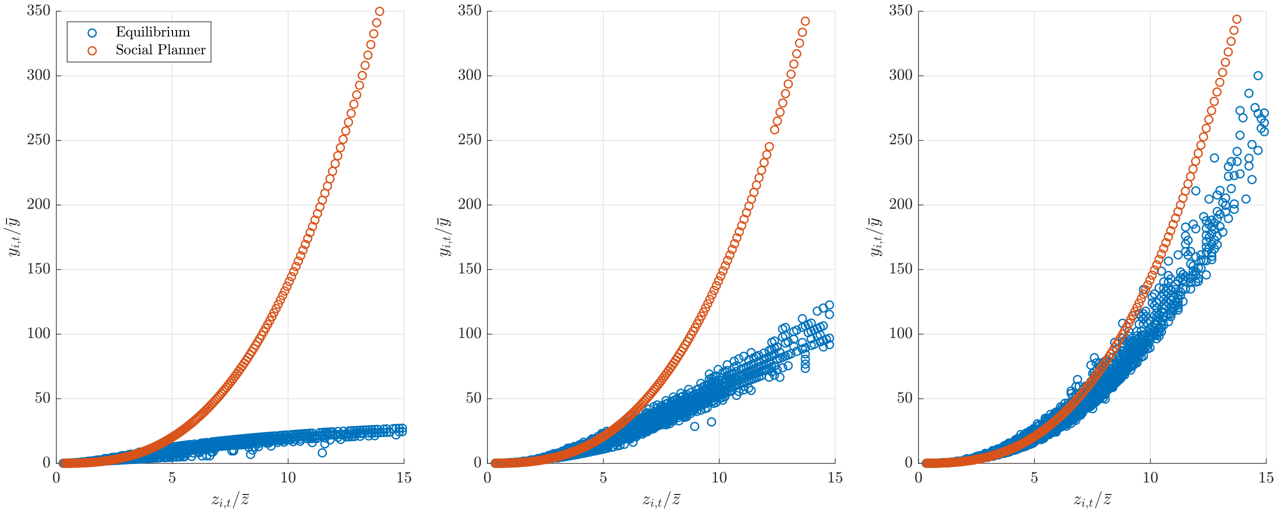
Aggregate Labor Supply Two forces work in opposite directions in affecting the differences of aggregate labor supply between the efficient and equilibrium allocations. On one hand, the more selective policy of the planner for entry and exit reduces the amount of labor required for financing the overhead costs of operating firms. On the other hand, labor is underutilized for the purposes of production in the equilibrium due to the 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 only goes up by 2.1% as it is mitigated by the lower use of labor in financing the entry cost of firms.

[26] 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 $Y = [\int_0^N y_i^{\theta^{-1}} di]^{\theta}$. For given amount of aggregate labor, L, every firms gets to produce $y_i = (L/N)^{\alpha}$ and the aggregate output is given by $Y = N^{\theta-\alpha}L$. Now if we shut down love for variety ($\theta = 1$), productivity is $N^{1-\alpha}$ indicating higher returns to scale with larger 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. Finally, if we shut down both channels, aggregate productivity is independent of N.

It is also worth mentioning that while aggregate production labor goes up for all three values of ϕ in Table 10, it is not always the case that *aggregate* labor supply is higher under the efficient allocation. For the counterfactual value of $\phi = 0.25$, there is so much less entry under the efficient allocation that aggregate labor falls by 5.0% while the production labor goes up by 5.3%.

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 depends 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 8 shows the scatter plot of firms' productivity and output for both the equilibrium and social planner's allocation for three different values of ϕ (low, calibrated and high). The figure show that it is indeed the case that with a larger ϕ the equilibrium allocation of customers is closer to that of the planner.

Figure 8: Comparison Efficient Allocation of Customers



(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 as they face higher marginal costs. Hence, aggregate TFP, output and concentration increase, but aggregate markups decrease and the economy gets closer to the efficient allocation.

7 Discussion

Before concluding our paper, here we discuss the implications of our model for the measurement of markups and our main model assumptions.

Implications for Measurement of Markups Our last fact in the empirical section is based on the assumption that markups are proportional to the Sales-to-COGS ratio at the firm level. There has been a recent debate on the validity of measuring markups through this method if SGA expenses are variable (see Traina, 2019; De Loecker et al., 2020). However, even if SGA is fully variable, an additional factor that matters for the measurement of markups is whether those costs are associated with the production process of the firm.

In particular, in our model, SGA expenses are semi-variable but the measurement of markups through the Sales-to-COGS ratio is still valid (as shown in Equation (3.12)). This is because of the fact that advertisement, while using labor as an input, does not contribute to the production process of the firm and it just increases demand. All production costs that are relevant for markups measurement are variable costs; however, in our model not all variable costs are production costs. Therefore, our measurement of markups in Section 2, that is based on the methodology of Hall (1988) and De Loecker and Warzynsik (2012) is appropriate.

Mechanisms for Customer Acquisition One of our main assumptions in the model is that customer acquisition at the firm level is independent of firm's price (as in, e.g., Arkolakis, 2010; Drozd and Nosal, 2012; Sedláček and Sterk, 2017). In particular, in our model customer acquisition can have the interpretation of a process through which potential customers become "aware" of the product or firm (as modeled by Perla, 2019). Once they are aware of the product, the price affects their demand only in the intensive margin and it does not affect their decision to leave the firm. This is in contrast with a branch of the literature that models customer acquisition through dynamic pricing. Our departure from this literature is mainly motivated by our third documented fact that markups are not correlated with the size of firms' customer bases conditional on sales per customer. Fitzgerald et al. (2016) also present evidence for this mechanism. They 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 quantities grow.

The Relative Nature of Marketing In our model, we have made the assumption that advertising is relative (as in Drozd and Nosal, 2012). This assumption implies that advertising does not increase the total number of matches generated within an industry and only affects the allocation of customers across firms. There are two aspects to this assumption. The first aspect is that the total number of customers available to an industry is fixed. This ensures that independent of how much firms spend on advertising within an industry they cannot create new customers. We think this is a reasonable assumption because there is only a fixed number of potential individuals within the economy that can be matched to products (see e.g. Arkolakis, 2010; Perla, 2019).

Moreover, a second aspect of our approach is that conditional on this constraint, all customers that are available are matches to one firm. This is in contrast to search models which assume that some agents can remain unmatched in the labor market, which is underneath the notion of unemployment in those models. However, we believe in the case of product markets, our assumption is reasonable in the sense that everyone who desires a product will find at least one seller that produces it.

Finally, the assumption that consumers are only matched to a single firm is not very restrictive. For instance, in the Nielsen database, within a product group-year, the median household consumes products produced only by one firm. This assumption becomes even more appropriate for durable goods. Furthermore, from a pure modeling perspective, even if consumers are matched to several firms, one can treat every match as a separate member and instead increase the measure of household members. However, this quantity is normalized to a unit measure in our model, which renders this assumption to be without loss of generality in our setup.[27]

8 Conclusion

Recent evidence suggests that concentration, market power and profit rates have increased in the U.S. over the past few decades. In this paper, we revisit the role of the extensive and intensive margins of demand in firms' market share and market power. Using a novel dataset that merges information from the consumer and producer side, we document that while firms' sales grow mainly through acquiring more customers, their market power is only correlated with their sales per customer. Moreover, we find that firms' non-production

[27] It is also important to note that in our characterization of the efficient allocation, even the planner is subject to these restrictions: she cannot create new customers and she cannot match one customer to more than one firm. 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.

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 micro-founded 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, in our model a certain degree of market power is desirable, as it compensates more productive firms for their investment in customer acquisition and improves the allocation of demand. Moreover, higher profits of larger firms are partly due to the returns on their past investment in their customer base. 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.

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.

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.

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 and Frederic Warzynsik, "Markups and Firm-Level Export Status," *American economic review*, 2012, *102* (6), 2437–71.

_ , 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.

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.

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.

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

Hall, Robert E, "The Relation Between Price and Marginal Cost in US Industry," *Journal of Political Economy*, 1988, *96* (5), 921–947.

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.

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, M, "Heterogeneous Markups, Growth and Endogenous Misallocation," 2019.

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.

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

A.1 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.

A.2 Compustat Variables

This section describes the cleaning process to generate the Compustat dataset used in this paper. The cleaning of Compustat follows the previous studies that use the same data to analyze the cost of production over time (De Loecker et al. 2020; Traina 2019).

We download and construct the following variables from Compustat:

- *Global company key* (mnemonic `gvkey`): Compustat's firm id.
- *Cusip*: identifier for "a specific security issue of a company."
- *Year* (mnemonic `fyear`): the fiscal year.
- *ISO country code, incorporation* (mnemonic `FIC`): indicates the country in which a company was incorporated.
- *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."
- *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."
- *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.
- *Assets, total* (mnemonic AT).
- *Acquisitions* (mnemonic AQC): this variable constitutes the costs relating to aqcuisition of another firm.
- *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.
- *interest rate on capital, r_t*: we define $r_t = (i_t - \pi_t) + \delta$, where i_t is the effective federal funds rate downloaded from FRED, π_t is the inflation (consumer prices) downloaded from FRED and δ is the depreciation rate, which we set at 12 %.
- *Advertising* (mnemonic XAD): this variable contains the cost of media advertising and promotional expenses.
- *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, XSGA, XAD and AT.

A.3 Compustat Data Cleaning

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

The raw dataset contains 436,891 observations for 33,327 firms. Next, 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.

Our final dataset contains 242,155 observations for 20,252 firms.

Combine Computstat, BEA IO table, and Worldscope data We combine the Compustat data with the BEA input-output data and the Worldscope data by taking the following steps, in order:

1. We merged NAICS codes to each of the IOCode industries, as provided by the BEA.
2. We improved the number of missing NAICS codes in Compustat by using a concordance table from the SIC industry codes to NAICS (our data is not missing SIC codes).
3. We merged the input-output data to Compustat using the NAICS code in a best merge case scenario, i.e. we first attempt to merge it on the 6-digit industries, for those that fail to merge we attempt the merge on the 5-digit industries, etc.
4. Finally, we merged the Compustat to WorldScope using the Cusip id.

A.4 SGA and COGS components

We document the components of SGA and COGS.

Selling, General and Administrative Expense (SGA): According to Compustat, SG&A "represents 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." The following expenses are allocated as components of SGA:

- Accounting expense
- Advertising expense
- Amortization of research and development costs
- Bad debt expense (provision for doubtful accounts)

- Commissions
- Corporate expense
- Delivery expenses
- Directors' fees and remuneration
- Engineering expense
- Extractive industries' lease rentals or expense, delay rentals, exploration expense, research and development expense, and geological and geophysical expenses, drilling program marketing expenses, and carrying charges on nonproducing properties
- Financial service industries' labor, occupancy and equipment, and related expenses
- Foreign currency adjustments when included by the company
- Freight-out expense
- Indirect costs when a separate Cost of Goods Sold figure is given
- Labor and related expenses (including salary, pension, retirement, profit sharing, provision for bonus and stock options, employee insurance, and other employee benefits when reported below a gross profit figure)
- Legal expense
- Marketing expense
- Operating expenses when a separate Cost of Goods Sold figure is given and no Selling, General, and Administrative Expense figure is reported
- Parent company charges for administrative services
- Recovery of allowance for losses
- Research and development companies' company-sponsored research and development
- Research and development expense
- Research revenue that is less that 50% of total revenues for 2 years
- Restaurants' preopening and closing costs
- Retail companies' preopening and closing costs and rent expense
- Severance pay (when reported as a component of Selling, General and Administrative Expenses)
- State income tax when included by the company
- Strike expense
- Stock-based compensation when reported below a gross profit figure

Costs of Goods Sold (COGS): According to Compustat, COGS "represents 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." Some firms report a breakdown of COGS, while others do not. For nonmanufacturing companies that do not report a breakdown, the total operating costs is considered as the COGS. However, in the case that a breakdown is provided, then the following expenses are considered components of COGS:

- Agricultural, aircraft, automotive, radio and television manufacturers' amortization of tools and dies
- Airlines' mutual aid agreements
- Amortization of deferred costs (i.e., start-up costs)
- Amortization of software costs and amortization of capitalized software costs
- Amortization of tools and dies where the useful life is two years or less
- Banks' interest expense on deposit
- Cooperatives' patronage dividends
- Customer-sponsored research and development expense for research and development companies
- Departmental costs
- Direct costs (when a separate Selling, General, and Administrative figure is reported)
- Direct labor
- Distribution and editorial expenses
- Expenses associated with sales-related income from software development
- Expenses of equity method joint ventures if reported as operating expenses
- Extractive industries' lease and mineral rights charged off and development costs written off
- Freight-in
- Heat, light, and power
- Improvements to leased property
- Insurance and safety
- Labor and related expenses reported above a gross profit figure (including salary, pension, retirement, profit sharing, provision for bonus and stock options, and other employee benefits)
- Land developers' investment real estate expense
- Lease expense
- Licenses
- Maintenance and repairs
- Operating expenses
- Real estate investment trusts' advisory fees
- Reimbursement for out of pocket expenses when reported as part of Cost of Goods Sold on the Income Statement

- Rent and royalty expense
- Restaurants' franchise fees
- Salary expense
- Stock based compensation when no gross profit figure is reported or when it is reported above a gross profit figure
- Supplies
- Taxes other than income taxes
- Terminals and traffic
- Transportation
- Warehouse expense
- Write-downs of oil and gas properties

Note that SGA and COGS share a few components; Compustat allocate them according to how firms report these variables.

B Additional Empirical Results

B.1 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:

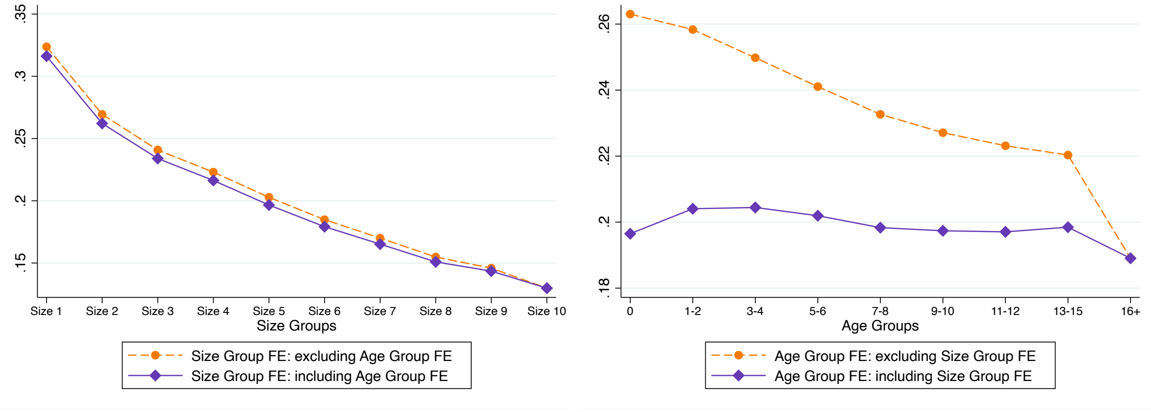
$$y_{it} = \sum_{m=1}^{10} \alpha_m \mathbb{1}(\text{size group}_{it} = m) + \sum_{a=1}^{9} \alpha_a \mathbb{1}(\text{age group}_{it} = a) + \alpha_s + \alpha_t + \varepsilon_{it}$$

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).[28] The sector and year fixed effects are included to analyze the average SGA-to-OPEX within each sector and each year.

We find that small firms have substantially larger non-production expenses compared to large firms, as presented in Figure C.1. Figure C.1a shows that smallest firms spend more than 30% of their expenses on non-production, while the largest firms spend approximately

[28] 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.

Figure C.1: SGA Share of OPEX: Size vs. Age



(a) Size Profile

(b) Age Profile

Notes: Figure C.1a plots the size profiles of the COGS-to-OPEX ratio (α_m) and Figure C.1b plots the age profiles of the SGA-to-OPEX ratio (α_a) by estimating equation B.1. 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 C.1a, 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 C.1b, 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.

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 C.1b shows that there is a negligible difference in the share of non-production expenses across old and young firms.

B.2 Robustness

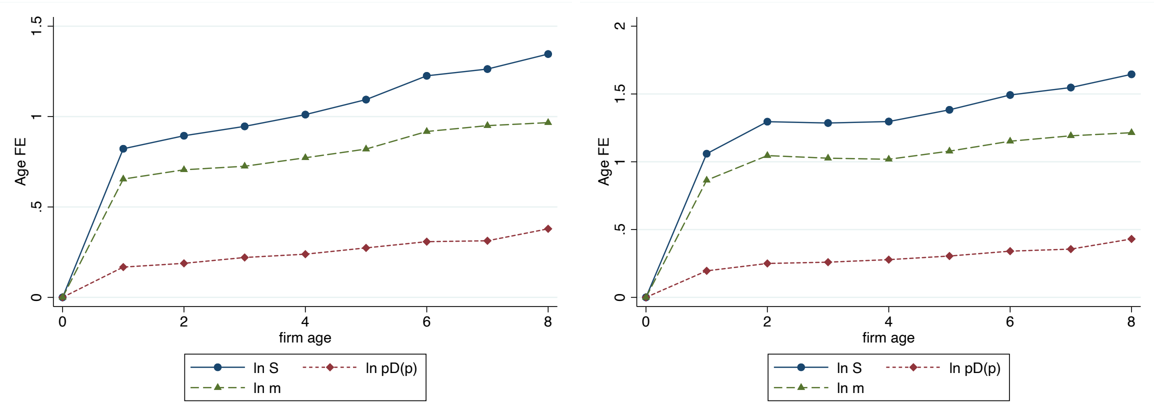
This section document additional results to confirm our main empirical findings with different specifications.

B.2.1 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 C.2 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.

Figure C.2: Decomposition of Firm Sales Growth by Firm Age



(a) Survive 3 Consecutive Years

(b) Survive 5 Consecutive Years

Notes: Figures C.2a and Figure C.2b 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 C.2a and 19,603 number of observations and 2,997 firm used in Figure C.2b.

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 C.3. 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 customers in explaining sales remain the same, explaining approximately 70% of sales on average.

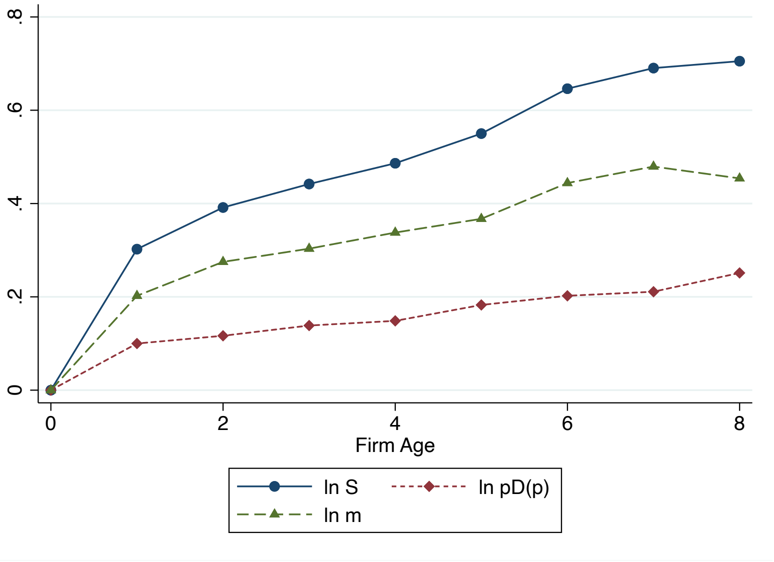
B.2.2 Semi-variable Nature of SGA

Table C.2 presents the regression results that correspond to Figure 2. 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 C.4 replicates Figure 2 using the full sample. Still, the SGA is more variable than COGS and less variable than SGA.

Figure C.5 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.

Figure C.3: Decomposition of Firm Sales Growth by Firm Age



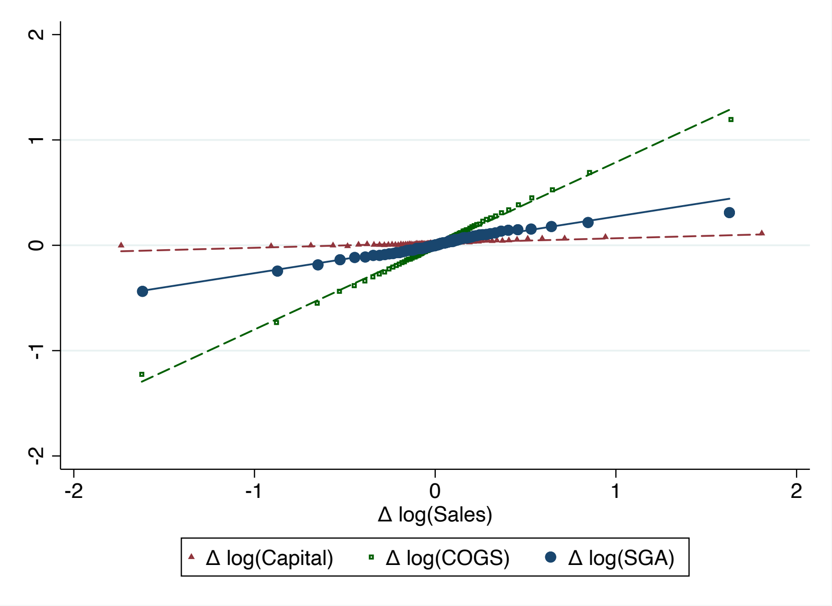
Notes: Figure C.3 replicates Figure 1 by using average monthly sales per firm and year.

Table C.2: The Semi-variable Nature of SGA

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	$\Delta ln(COGS)$	$\Delta ln(COGS)$	$\Delta ln(SGA)$	$\Delta ln(SGA)$	$\Delta ln(Capital)$	$\Delta ln(Capital)$	$\Delta ln(R\&D)$	$\Delta ln(R\&D)$
$\Delta log(Sales)$	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

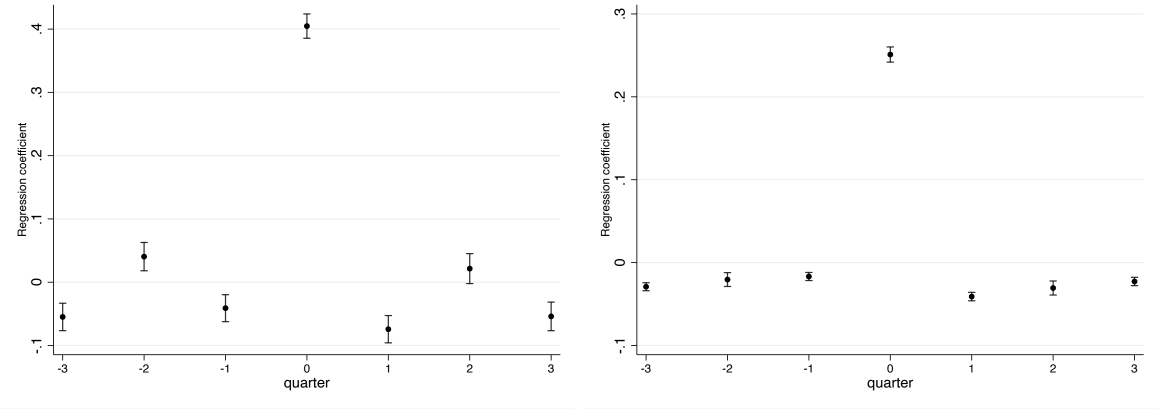
Notes: The dependent variables are the quarterly change in log COGS, SGA and Capital. The estimation method used in all columns is OLS. Standard errors (in parentheses) are clustered at the firm level.

Figure C.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 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 C.5: Cross-correlation



(a) Trimmed Sample

(b) Full Sample

Notes: Figure C.5a uses the trimmed sample presented in the main body of the paper, and Figure C.5b uses the full sample. 95% confidence intervals are presented for every estimate. Figure C.5a replicates the column (4) in Table C.2 at quarter = 0.

B.2.3 Markups, Sales per customers, and Sales

Table C.3 replicate Table 4 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. Since there are firms that generate a minimal amount of sales in the Nielsen data, Table C.4 considers the same analyses by excluding observations that have annual firm sales of less than 1 million USD. For the robustness checks, Table C.5 and C.6 consider the sample that has more than 2 and 0.5 million USD, respectively.

Regardless of controlling sales and using different samples, our results still show a strong correlation between markups and sales per customer. Consistent with previous studies, we observe a positive correlation between markups and sales, but this correlation is weaker than the one between markups and sales per customer. Our analysis suggests the importance of the sales per customer in understanding the firm-level markups.

Table C.3: Markups, Nielsen Sales, and Nielsen Sales per Customer

	(1)	(2)	(3)	(4)	(5)
ln pD(p)_{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.

Table C.4: Markups, Nielsen Sales, and Nielsen Sales per Customer, > 1 million sales

	(1)	(2)	(3)	(4)	(5)
ln $pD(p)_{it}$	0.142^{***}	0.143^{***}	0.103^{***}	0.104^{***}	0.105^{***}
	(0.043)	(0.044)	(0.029)	(0.029)	(0.033)
ln S_{it}	0.017^{*}	0.017^{*}	0.019^{**}	0.020^{**}	0.020^{**}
	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)
Observations	1741	1741	1741	1741	1741
R^2	0.104	0.106	0.422	0.423	0.449
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. Only firm-year bins that have more than one million sales are used in the analyses to exclude exceptionally small firms.

Table C.5: Markups, Nielsen Sales, and Nielsen Sales per Customer, > 2 million sales

	(1)	(2)	(3)	(4)	(5)
ln $pD(p)_{it}$	0.143^{***}	0.147^{***}	0.099^{***}	0.101^{***}	0.105^{***}
	(0.047)	(0.048)	(0.032)	(0.032)	(0.036)
ln S_{it}	0.023^{**}	0.023^{**}	0.025^{***}	0.025^{***}	0.026^{***}
	(0.010)	(0.010)	(0.009)	(0.009)	(0.010)
Observations	1600	1600	1600	1600	1600
R^2	0.108	0.112	0.443	0.445	0.469
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. Only firm-year bins that have more than two million sales are used in the analyses to exclude exceptionally small firms.

Table C.6: Markups, Nielsen Sales, and Nielsen Sales per Customer, $>$ .5 million sales

	(1)	(2)	(3)	(4)	(5)
ln $\mathrm{pD(p)_{it}}$	0.136***	0.136***	0.094***	0.093***	0.094***
	(0.040)	(0.041)	(0.026)	(0.027)	(0.029)
ln $\mathrm{S_{it}}$	0.016*	0.016*	0.017**	0.017**	0.017**
	(0.008)	(0.008)	(0.007)	(0.007)	(0.008)
Observations	1872	1872	1872	1872	1872
R^2	0.099	0.100	0.414	0.415	0.437
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. Only firm-year bins that have more than one-half million sales are used in the analyses to exclude exceptionally small firms.

C Derivations and Proofs

C.1 Lemma 1

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

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

where

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

Where the last equality in Equation (C.2) follows from the expression of demand per match in Equation (3.4). Differentiating Equation (C.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. It follows that

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

□

C.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 with Equation (3.14), we have

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

□

C.3 Corollary 1

Proof. We show this 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 we have only used the elasticity of markup 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. □

C.4 Proposition 2

Proof. This relationship is obtained directly from the first order condition of the firms' 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 $l_{i,s,t}$. So 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 profit 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

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

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

C.5 Planner's Problem

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

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

subject to

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

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

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

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

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

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

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

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

Here Equation (C.4) requires that for every firm, their supply meets their allocated demand, Equation (C.5) is the Kimball aggregator that implicitly defines C_t given the planner allocation of demand, Equation (C.6) requires that labor supply meets demand for labor from production, advertisement and overhead costs, Equation (C.7) requires that the matching market clears, Equation (C.8) is the law of motion for the set of operating firms given an entry/exit policy by the planner, Equation (C.9) determines firm i's evolution of matches given an allocation for marketing, Equation (C.10) requires the non-negativity of labor for marketing and the constraint that while the planner can separate matches from firms the separation rate should be at least δ, and finally Equation (C.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 matches, which is also well-defined in the limit when $\bar{L}_{s,t} \to 0$. □

C.6 Lemma 2

Proof. We do this by construction. Suppose at any given time t, a choice for N_t is fixed. Suppose now 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. Now to verify that these values implement $\mathcal{A}$, observe that

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

□

C.7 Proposition 3

Proof. The results in this Proposition follow from the first order conditions of the planner's problem in Equation (C.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 i and t, let $\beta^t \eta_{c,i,t} di$ be the shadow cost on Equation (C.4); for t, let $\beta^t \eta_{Y,t}$, $\beta^t \eta_{L,t}$ and $\beta^t \eta_{m,t}$ be the shadow costs on Equation (C.5), Equation (C.6) and Equation (C.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 $j \in m_{i,t}$.

Now, the first order conditions with respect to $q_{i,j,t}$ are:

$$\eta_{c,i,t} C_t = Y'(q_{i,j,t})\eta_{Y,t}, \quad \forall j \in m_{i,t} \tag{C.12}$$

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

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

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

$$\eta_{c,i,t} q_{i,t} C_t + \eta_{m,t} = Y(q_{i,t})\eta_{Y,t} \tag{C.14}$$

Notice that in deriving this first order condition, we have ignored the constraint in Equation (C.9). The reason that we can do this goes back to Lemma 2 which states 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 (C.9) and then use Lemma 2 to show that it is satisfied.

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

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

Now, since $\eta_{\Upsilon,t} \neq 0$,[29] it follows that

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

Notice that the left hand side of this equation is only a function of $q_{i,t}$ and it is strictly monotonic in $q_{i,t} > 0$.[30] Moreover, the right hand side of the equation is only a function of time t shadow costs and is independent of i. Hence,

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

Replacing this last equation into Equation (C.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 for matches in Equation (C.7) and the last statement uses the strict monotonicity of $\Upsilon(x)$ and the fact that $\Upsilon(1) = 1$.

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

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

Divide this equation by the first order condition for $m_{i,t}$ in Equation (C.12) to get

$$z_{i,t}l_{i,p,t}^{\alpha-1} = C_t\frac{\eta_{L,t}}{\eta_{\Upsilon,t}}$$

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

$$\int_{j\in m_{i,t}} c_{i,j,t}dj = m_{i,t}C_t = z_{i,t}\left(\frac{C_t}{z_{i,t}}\frac{\eta_{L,t}}{\eta_{\Upsilon,t}}\right)^{\frac{\alpha}{\alpha-1}} \Rightarrow m_{i,t} = \left(\frac{z_{i,t}}{C_t}\right)^{\frac{1}{1-\alpha}}\left(\frac{\eta_{L,t}}{\eta_{\Upsilon,t}}\right)^{\frac{\alpha}{\alpha-1}}$$

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

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

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

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

□

C.8 Proposition 4

Proof. A few observations are useful in proving this Proposition. First, Lemma 2 allows us to ignore constraints in Equation (C.9), and Equation (C.10), solve for the optimal allocation of matches and use the results of the 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 $z_{i,t}$.

Now, to characterize this distribution, define $n_t(z)$ as the density of operating firms at time t that have productivity less than 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 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:

$$n_t(z) = \mathbf{1}_t(z) \int_{z_{-1}\in\mathbb{R}_+} f(z|z_{-1})(\lambda\Gamma(z_{-1}) + \nu n_{t-1}(z_{-1}))dz_{-1} \tag{C.15}$$

Here $f(z|z_{-1})$ denotes the conditional density of $z|z_{-1}$, which is governed by an AR(1) per Equation (3.7) and $\Gamma(.)$ is the CDF of the productivity distribution of the entrants from Equation (3.6).

Equation (C.15) merely shows that the density of firms at time t with productivity z comes either from entrants in the last period who transitioned to 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 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:

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

Now, 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 (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

$$\begin{aligned}
\max_{\{C_t,L_t,L_{p,t},\mathbf{1}_t(z)_{z\in\mathbb{R}_+}\}_{t\geq 0}} & \sum_{t=0}^{\infty}\beta^t\left[\frac{C_t^{1-\gamma}}{1-\gamma}-\xi\frac{L_t^{1+\psi}}{1+\psi}\right]\\
s.t.\quad & C_t = \left[\int_{z\in\mathbb{R}_+} z^{\frac{1}{1-\alpha}} n_t(z)dz\right]^{1-\alpha} L_{p,t}^{\alpha}\\
& L_t = \chi\int_{z\in\mathbb{R}_+} n_t(z)dz + L_{p,t} + \bar{L}_{s,t}\\
& n_t(z) = \mathbf{1}_t(z)\int_{z_{-1}\in\mathbb{R}_+} f(z|z_{-1})(\lambda\Gamma(z_{-1})+\nu n_{t-1}(z_{-1}))dz_{-1}, \forall z\in\mathbb{R}_+\\
& n_{-1}(z) \text{ given.}
\end{aligned}$$

Now, for any t let $\beta^t\eta_{C,t}$, $\beta^t\eta_{L,t}$ and $\beta^t\eta_{n,t}(z)$ be the Lagrange multipliers on the constraints in Equations (C.8), (C.8) and (C.8) respectively. The implied first order conditions for C_t, L_t and $L_{p,t}$ are

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

Combining these first order conditions, we arrive at the standard 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 $\mathbf{1}_{i,t}(z)=0$.[31] Now, note that conditional on keeping a type, the

[31] One could allow the planner to keep a fraction of firms with productivity z but since the measure of each

FOC with respect to $n_t(z) > 0$ is

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

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. Now, dividing this equation by $\eta_{C,t}$, replacing $\alpha C_t^* = W_t^* L_{p,t}^*$ and plugging in $W_t^* = \eta_{L,t}/\eta_{C,t}$ and $\eta_{C,t} = C_t^{*-\gamma}$ we arrive at:

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

While this characterizes the shadow cost/benefit of the firm, in units of consumption, conditional on being kept, 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 $G_t^*(z) \equiv \max_{\mathbf{1}_t(z)} \frac{\eta_{n,t}(z)}{\eta_{C,t}}\mathbf{1}_t(z)$, 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 forward, and re-writing this in sequential form for firm i gives us the expression in the proposition. □

C.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:

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

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

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

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 (3.21)

type is infinitesimal, assuming that the planner either keeps every type or lets all of that type go is without loss of generality.

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

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

□

C.10 Proposition 6

Proof. Recall 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

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

Now notice that

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

□

C.11 Proposition 7

Proof. This can be derived from combining Equations (C.10) and (C.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 $\alpha > \phi$. □

D 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 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.

D.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:

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

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:

$$
\begin{aligned}
\mathcal{N}_t(m, z') = & \int_{M \times Z} F(z'|z) \mathbb{1}_{\{m^*(m_{-1},z) \leq\}} \nu \mathbb{1}_{\{v(m_{-1},z) \geq 0\}} d\mathcal{N}(m_{-1}, z) \\
& + \lambda \int_{M \times Z} F(z'|z) \mathbb{1}_{\{m^*(0,z) \leq m\}} \mathbb{1}_{\{v(0,z) \geq 0\}} d\Gamma(z),
\end{aligned}
$$

where $F(z'|z)$ is the Markov chain given by the AR(1) productivity process, Γ 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 stochastic discount factor $U_{c,t+1}/U_{c,t} = 1$. 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:

$$Q_M(s, m'(s)) = \left[\mathbb{1}_{m'(s) \in [m_{j-1}, m_j]} \frac{m'(s) - m_j}{m_j - m_{j-1}} + \mathbb{1}_{m'(s) \in [m_j, m_{j+1}]} \frac{m_{j+1} - y'(s)}{m_{j+1} - m_j}\right] \tag{D.1}$$

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 (D.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

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

where Γ 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 $X = (\nu \mathbb{1}_{v(\boldsymbol{s}) \geq 0} \boldsymbol{\mathcal{N}} +$

$\lambda \mathbb{1}_{v(\boldsymbol{s})\geq 0}\Gamma)' x(\boldsymbol{s})$. 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 a go to **Step 2**.

D.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 $Y = 1$ and the wage $W = 1$ 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.

E Additional Model Analysis

E.1 Identification of Model Parameters

In this Section, we formally guide the discussion of the identification of model parameters. Panel A of Figure C.6 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 C.6 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.

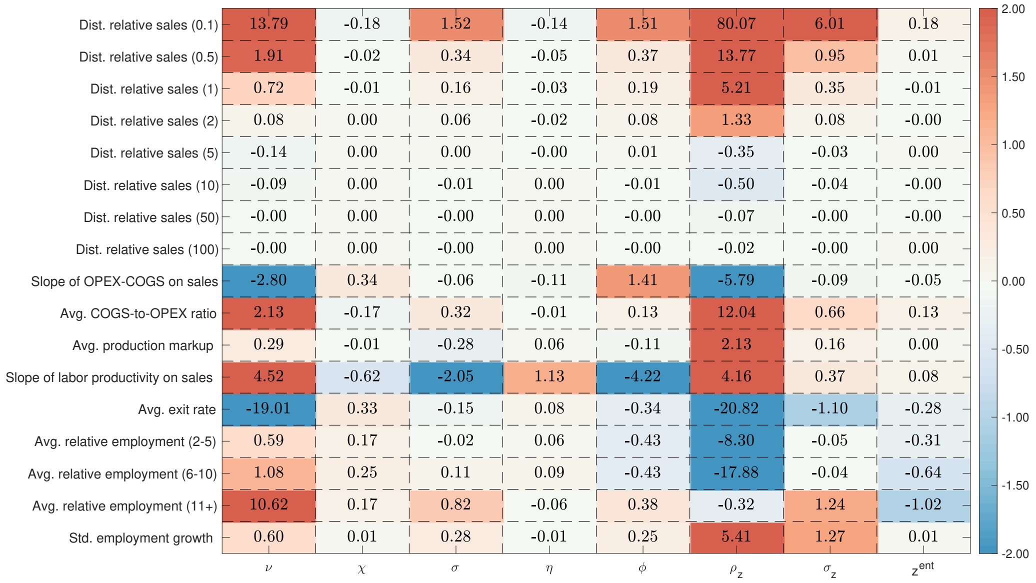
E.2 Additional Calibration Figures

Here we provide additional results about the goodness of fit of the calibrated model. Figure C.7 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 C.8 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 the same 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 C.9 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.

E.3 Additional Analysis of Model in Steady State

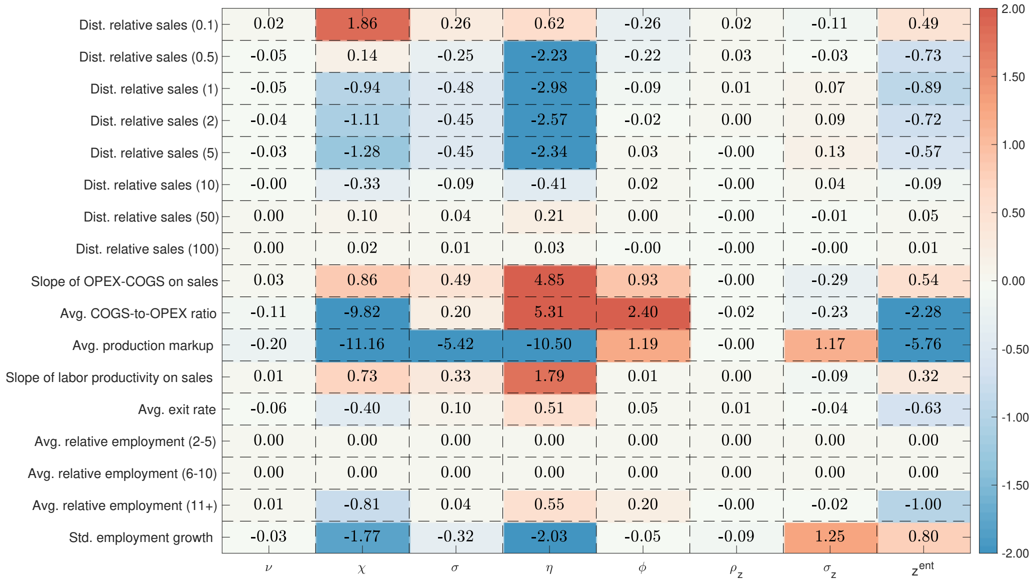
In this section, we further describe how the model works in steady state. First, we describe firms' optimal policies. Then, we show the average firm dynamics, taking selection into

Figure C.6: Parameter Identification



	ν	χ	σ	η	ϕ	ρ_z	σ_z	z^{ent}
Dist. relative sales (0.1)	13.79	-0.18	1.52	-0.14	1.51	80.07	6.01	0.18
Dist. relative sales (0.5)	1.91	-0.02	0.34	-0.05	0.37	13.77	0.95	0.01
Dist. relative sales (1)	0.72	-0.01	0.16	-0.03	0.19	5.21	0.35	-0.01
Dist. relative sales (2)	0.08	0.00	0.06	-0.02	0.08	1.33	0.08	-0.00
Dist. relative sales (5)	-0.14	0.00	0.00	-0.00	0.01	-0.35	-0.03	0.00
Dist. relative sales (10)	-0.09	0.00	-0.01	0.00	-0.01	-0.50	-0.04	-0.00
Dist. relative sales (50)	-0.00	0.00	-0.00	0.00	-0.00	-0.07	-0.00	-0.00
Dist. relative sales (100)	-0.00	0.00	-0.00	0.00	-0.00	-0.02	-0.00	0.00
Slope of OPEX-COGS on sales	-2.80	0.34	-0.06	-0.11	1.41	-5.79	-0.09	-0.05
Avg. COGS-to-OPEX ratio	2.13	-0.17	0.32	-0.01	0.13	12.04	0.66	0.13
Avg. production markup	0.29	-0.01	-0.28	0.06	-0.11	2.13	0.16	0.00
Slope of labor productivity on sales	4.52	-0.62	-2.05	1.13	-4.22	4.16	0.37	0.08
Avg. exit rate	-19.01	0.33	-0.15	0.08	-0.34	-20.82	-1.10	-0.28
Avg. relative employment (2-5)	0.59	0.17	-0.02	0.06	-0.43	-8.30	-0.05	-0.31
Avg. relative employment (6-10)	1.08	0.25	0.11	0.09	-0.43	-17.88	-0.04	-0.64
Avg. relative employment (11+)	10.62	0.17	0.82	-0.06	0.38	-0.32	1.24	-1.02
Std. employment growth	0.60	0.01	0.28	-0.01	0.25	5.41	1.27	0.01

(a) Sensitivity of Moments

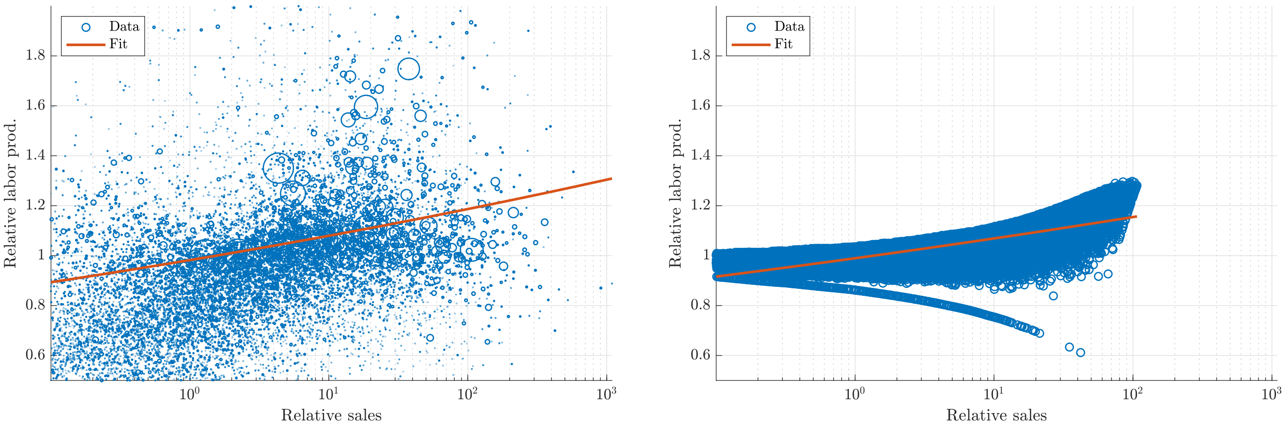


	ν	χ	σ	η	ϕ	ρ_z	σ_z	z^{ent}
Dist. relative sales (0.1)	0.02	1.86	0.26	0.62	-0.26	0.02	-0.11	0.49
Dist. relative sales (0.5)	-0.05	0.14	-0.25	-2.23	-0.22	0.03	-0.03	-0.73
Dist. relative sales (1)	-0.05	-0.94	-0.48	-2.98	-0.09	0.01	0.07	-0.89
Dist. relative sales (2)	-0.04	-1.11	-0.45	-2.57	-0.02	0.00	0.09	-0.72
Dist. relative sales (5)	-0.03	-1.28	-0.45	-2.34	0.03	-0.00	0.13	-0.57
Dist. relative sales (10)	-0.00	-0.33	-0.09	-0.41	0.02	-0.00	0.04	-0.09
Dist. relative sales (50)	0.00	0.10	0.04	0.21	0.00	-0.00	-0.01	0.05
Dist. relative sales (100)	0.00	0.02	0.01	0.03	-0.00	-0.00	-0.00	0.01
Slope of OPEX-COGS on sales	0.03	0.86	0.49	4.85	0.93	-0.00	-0.29	0.54
Avg. COGS-to-OPEX ratio	-0.11	-9.82	0.20	5.31	2.40	-0.02	-0.23	-2.28
Avg. production markup	-0.20	-11.16	-5.42	-10.50	1.19	-0.00	1.17	-5.76
Slope of labor productivity on sales	0.01	0.73	0.33	1.79	0.01	0.00	-0.09	0.32
Avg. exit rate	-0.06	-0.40	0.10	0.51	0.05	0.01	-0.04	-0.63
Avg. relative employment (2-5)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Avg. relative employment (6-10)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Avg. relative employment (11+)	0.01	-0.81	0.04	0.55	0.20	-0.00	-0.02	-1.00
Std. employment growth	-0.03	-1.77	-0.32	-2.03	-0.05	-0.09	1.25	0.80

(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.

Figure C.7: Model Fit: Labor Productivity and Sales

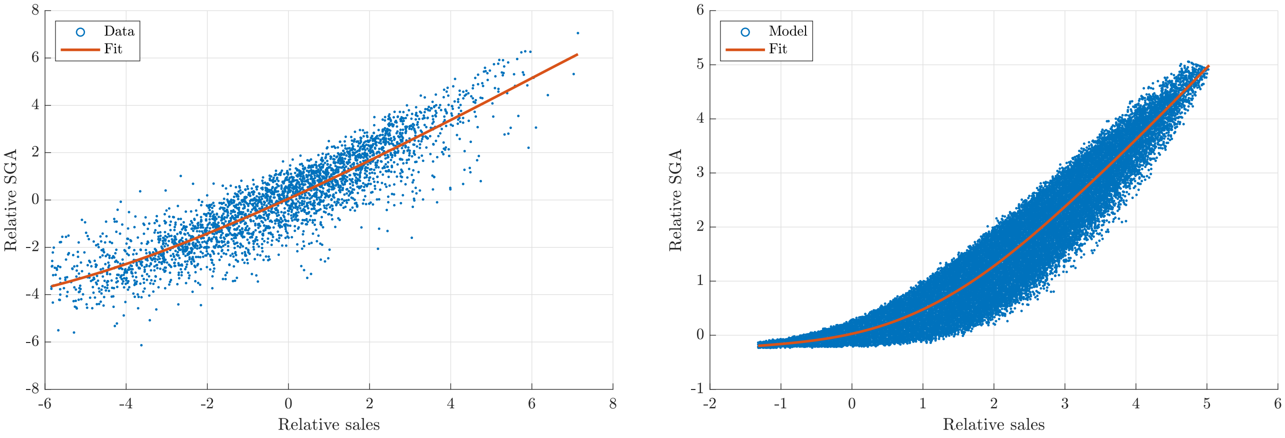


(a) Data

(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 linear best fit of the data. The x-axis is in log scale.

Figure C.8: Model Fit: SGA and Sales



(a) Data

(b) Model

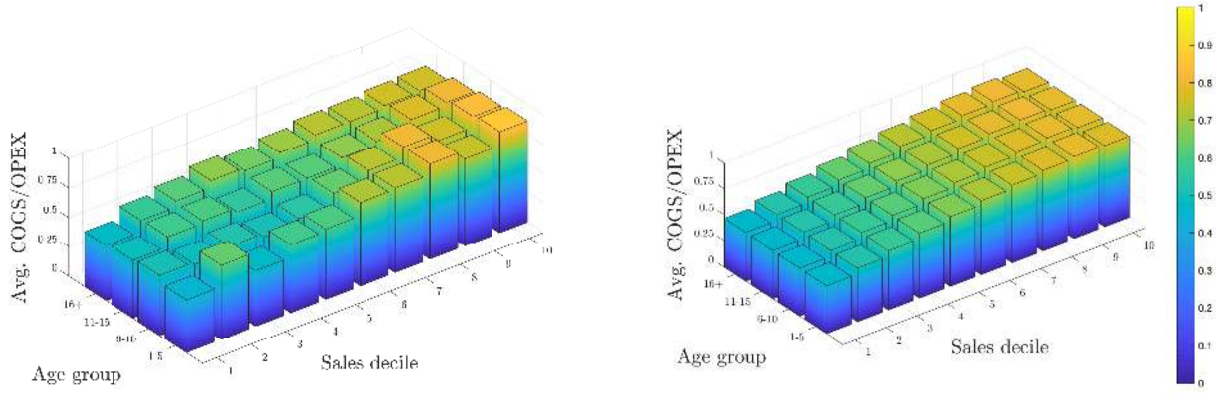
Notes: This figure plots the relationship between relative spending in SGA and relative sales. Panel A and B show the relationship obtained from the Compustat data and the model simulated data, respectively. The model data is obtained by simulating the model and restricting the sample to firms that are 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.

account.

Firms' Optimal Policies Figure C.10 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 relative customer base.

While optimal spending in $l_{i,s,t}$ decreases with the size of the customer base, production

Figure C.9: Steady-State COGS/OPEX by Size and Age



(a) Data

(b) Model

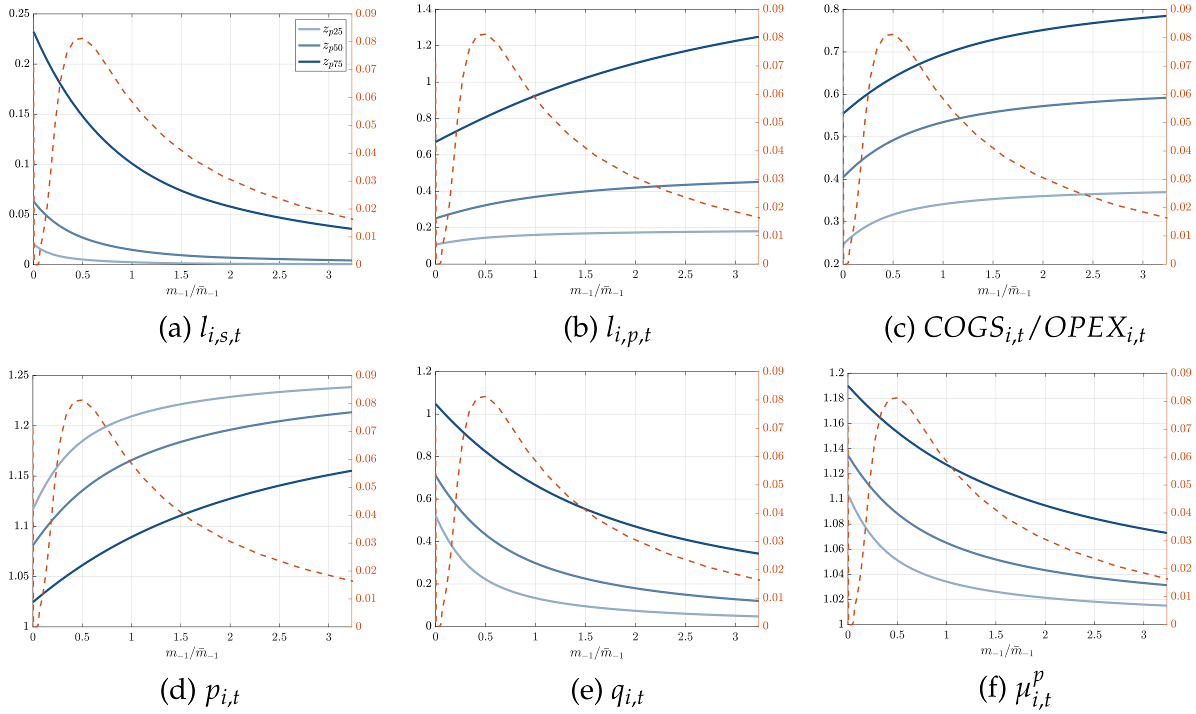
Notes: This figure plots the average COGS-to-OPEX ratio as a function of a firm's age and size. Panel A and B show the relationship obtained from the Compustat data and the model simulated data, respectively. The model data is obtained by simulating the model and restricting the sample to firms that 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.

labor is increasing in a firm's customer base. Thus, when firms have a small customer base, 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 C.11 plots firms' optimal exit policy 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. Despite this selection effect, 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.

Average Firm Dynamics with Shocks Figure C.12 plots the average firm dynamics taking selection into account. To construct this figure, we simulate a cohort of firms that starts with zero customer base and draws productivities from the distribution of entrants. As firms are

Figure C.10: Firms' Optimal Policies



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

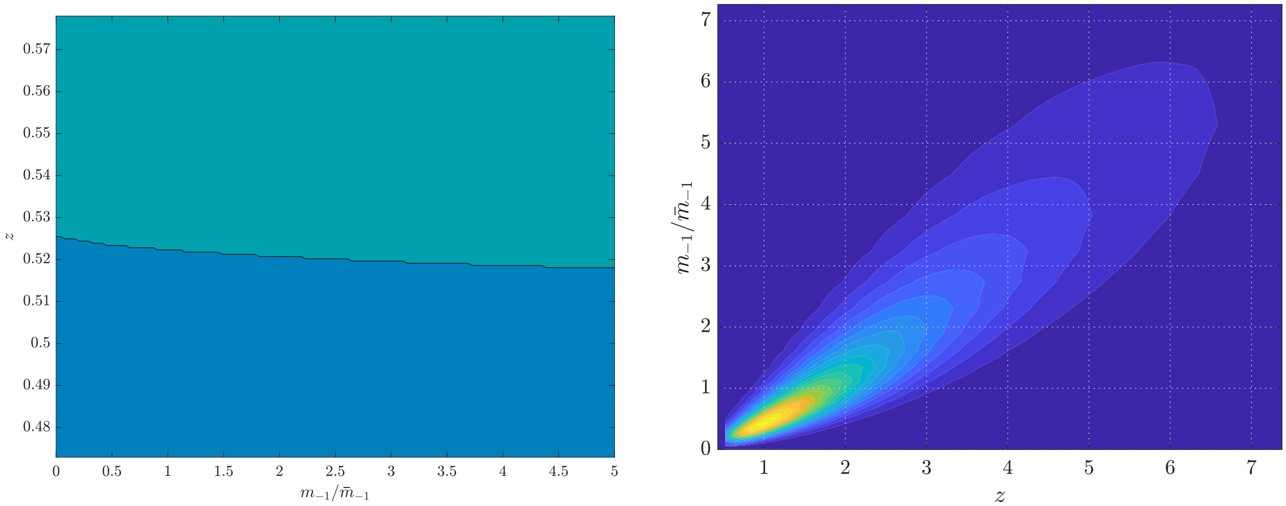
(d) $p_{i,t}$ (e) $q_{i,t}$ (f) $\mu^p_{i,t}$

Notes: These figures plot firms' policy functions. 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 relative customer base.

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 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 C.11: Optimal Exit and Stationary Distribution

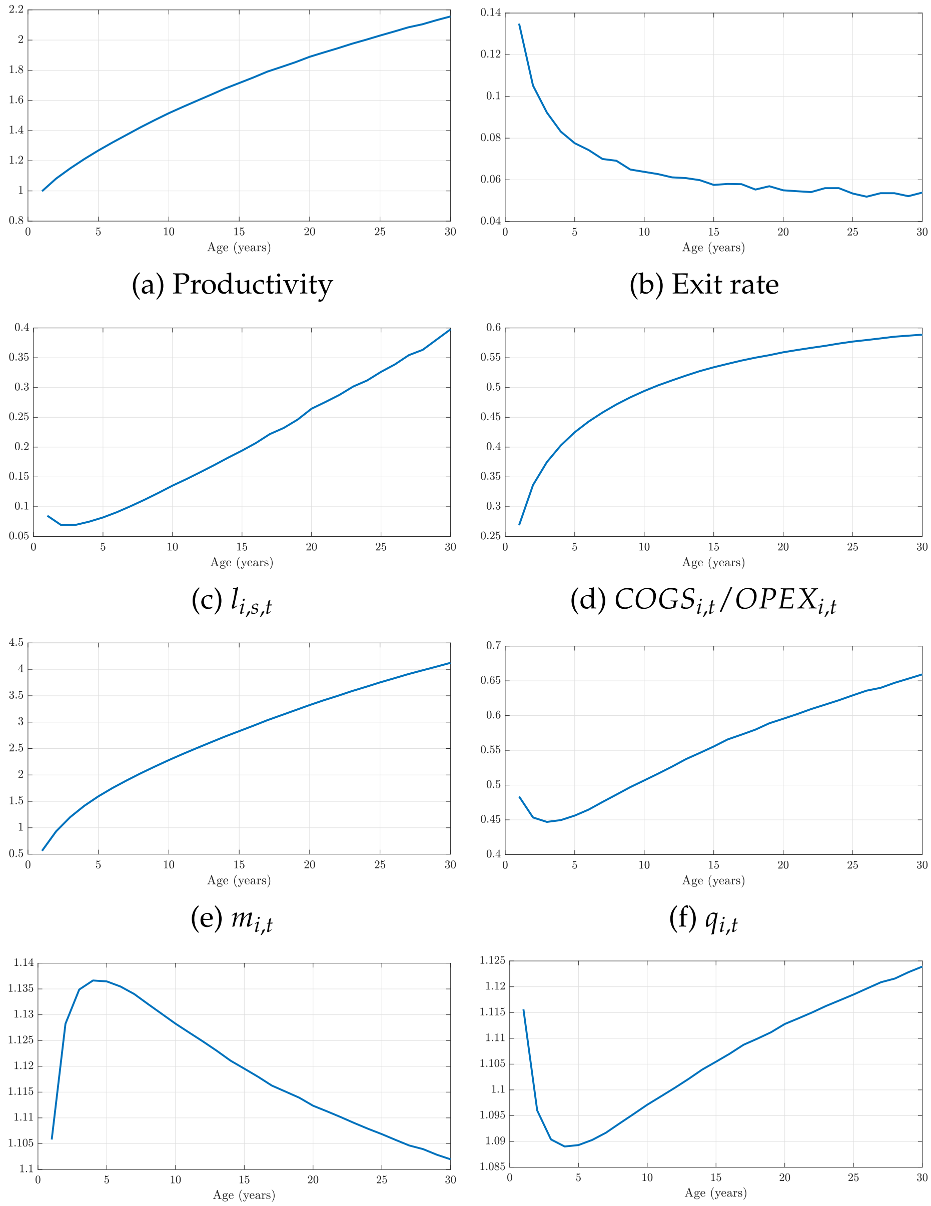


(a) Optimal exit policy

(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.

Figure C.12: 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.