NBER WORKING PAPER SERIES

CONCENTRATION, MARKET POWER, AND MISALLOCATION: THE ROLE OF ENDOGENOUS CUSTOMER ACQUISITION

Hassan Afrouzi
Andres Drenik
Ryan Kim

Working Paper 31415
http://www.nber.org/papers/w31415

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
June 2023

We thank David Argente, Ariel Burstein, Doireann Fitzgerald, and seminar participants at various institutions and conferences for valuable comments and suggestions. Luigi Caloi provided superb research assistance. Previous versions of this manurscript were circulated under the title "Growing by the Masses: Revisiting the Link between Firm Size and Market Power." Researcher(s)' own analyses calculated (or derived) based in part on data from Nielsen Consumer LLC and marketing databases provided through the NielsenIQ Datasets at the Kilts Center for Marketing Data Center at The University of Chicago Booth School of Business. The conclusions drawn from the NielsenIQ data are those of the researcher(s) and do not reflect the views of NielsenIQ. NielsenIQ is not responsible for, had no role in, and was not involved in analyzing and preparing the results reported herein. Hassan Afrouzi gratefully acknowledges support from Lenfest Junior Development Grant from Columbia University. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Concentration, Market Power, and Misallocation: The Role of Endogenous Customer Acquisition
Hassan Afrouzi, Andres Drenik, and Ryan Kim
NBER Working Paper No. 31415
June 2023
JEL No. D24,D43,D61,E22

ABSTRACT

This paper explores how different margins of market share are related to markups. Using merged microdata on producers and consumers, we document that a firm's market share is mainly related to its number of customers, while its price-cost markup is associated only with its average sales per customer. We develop a new model that reflects this empirical evidence and the endogenous nature of customer acquisition. When calibrated, this model predicts a higher degree of markup dispersion, which suggests greater efficiency losses due to customer misallocation. An analysis of the efficient allocation in this model reveals that compared with the equilibrium, aggregate TFP and output are 10.8% and 14% higher, respectively.

Hassan Afrouzi
Department of Economics
Columbia University
420 W 118th Street
New York, NY 10027
and NBER
ha2475@columbia.edu

Ryan Kim
School of Advanced International Studies
Johns Hopkins University
1717 Massachusetts Ave NW
Washington, DC 20036
rkim59@jhu.edu

Andres Drenik
Department of Economics
University of Texas at Austin
2225 Speedway
Austin, TX 78712
andres.drenik@austin.utexas.edu

An online appendix is available at: http://www.nber.org/data-appendix/w31415

1 Introduction

In standard macroeconomic models of endogenous markups—e.g., Atkeson and Burstein (2008), Klenow and Willis (2016)—firms with larger market shares charge higher price-cost markups. At a macro level, such mechanisms generate a close connection between market share distribution and markup dispersion, which is a source of resource misallocation due to firm-level wedges (Restuccia and Rogerson, 2008, Hsieh and Klenow, 2009). However, whereas in these models market share is driven by firms' sales to a representative household (the intensive margin of demand), in practice a firm's market share is affected by more than one margin. Specifically, firms spend a vast amount of resources on expanding their customer bases (the extensive margin of demand).[1] This leads us to ask: Do both the intensive and extensive margins of market share have the same relationship with a firm's market power (markups)? If not, how does this change our understanding of the ties between concentration, market power, and misallocation?

In this paper, we investigate how these two demand margins are related to firms' market shares and markups. Merging individual product-level consumption data and producer-level data, we find that firms' markups correlate *only* with their average sales per customer (the intensive margin), but not with the size of their customer base (the extensive margin). Yet only about 22% of the variation in firms' market share is tied to the margin associated with markups. To explore the macroeconomic implications of these findings, we develop a model with endogenous customer acquisition that is consistent with these facts. The model predicts an equilibrium relationship between market share and market power as in standard models; however, when calibrated to match the same distribution of firm size, it generates noticeably higher variation in markups across firms. This higher markup dispersion hints at higher efficiency losses due to misallocation of demand. To measure these losses rigorously, we characterize and quantify the first best allocation in our model: Relative to the efficient allocation, equilibrium aggregate TFP and output are 10.8% and 14% lower, respectively.

Our first contribution is to empirically investigate the relationship between different margins of market share and markups by merging the Nielsen Homescan Panel and Compustat datasets. First, our most novel finding is that firms' markups are not correlated with the size of their customer bases. Instead, they are positively associated *only* with their average sales per customer. Second, we find that around three-quarters of the variation across firms' market shares is explained by the margin that is not correlated with markups;

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

i.e., the extensive margin. This is consistent, both qualitatively and quantitatively, with the contemporary findings in Einav, Klenow, Levin, and Murciano-Goroff (2022) and Argente, Fitzgerald, Moreira, and Priolo (2021), who, using different data, show that most of the variation in firms' sales is driven by the size of their customer base. Third, we find that firms' non-production expenses are positively associated with the acquisition of new customers but not with their average sales per customer, which suggests that firms engage in activities in order to expand their customer bases by acquiring new customers.

These empirical findings inform the theoretical model we develop to understand the links between varying sources of market share and market power. In our model, at each period, a set of new firms draw different productivity levels and decide whether to enter the economy subject to fixed operating costs. Conditional on entry, these firms are monopolistically competitive and can spend resources to acquire new customers. These firms face semi-kinked demand curves a la Kimball (1995) from each customer, which indicates that each customer's demand is more elastic when a firm's relative price is larger (also known as Marshall's second law of demand). Therefore, whereas a firm's total customer base merely shifts its demand, as in Phelps and Winter (1970), the elasticity of this demand is determined by each customer's individual demand curve. Thus, the model generates a comovement between markups and average sales per customer, but not between markups and the number of customers (consistent with our first fact). Moreover, it allows firms to grow through both margins of demand (second fact), which implies an endogenous relationship between sales and non-production costs (third fact).

By allowing firms to grow through the extensive margin, our model breaks the *direct* relationship between market share and demand elasticity generated by the exogenous shape of the demand curve in standard models. Since market share in our model is determined through two separate margins—but markups are only correlated with one of those margins—conditional on each size group, there is a whole range of markups that firms within that group charge. However, on average, our model still creates a positive correlation between markups and market shares as an *equilibrium outcome* bearing unique counterfactual implications. Although the extensive margin does not correlate with markups conditional on sales per customer, this channel allows for a relationship between size and market power through the costs and benefits of customer acquisition: More productive firms expect to charge higher markups to their customers, anticipating higher gains from additional customers. Therefore, higher-markup firms also invest more in their customer bases. Hence, the model is consistent with a positive relationship

between market shares and markups, as in conventional models, but has notably different macroeconomic implications.

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

With the calibrated model at hand, we ask how does the model change our understanding of the relationship between market share and market power? To answer this question, we compare the equilibrium allocation with the one obtained in a version of the model that corresponds to a specification of conventional models that is *recalibrated* to match the same moments, including the size distribution of firms.[2] Comparing the two, we find that our model associates the same moments with a higher degree of markup dispersion, which anticipates a higher degree of welfare loss due to misallocation.

Motivated by this higher markup dispersion, we characterize and quantify the efficient allocation in our model. Under this allocation, the social planner increases aggregate productivity by allocating more customers to more productive firms while equalizing the relative demand per customer across weakly substitutable varieties. This result contrasts with the efficient allocation in conventional models, in which the planner uses the intensive margin of demand to target two mutually exclusive objectives: concentrate demand among more productive firms to increase aggregate productivity versus equalize demand across varieties to eliminate utility losses from demand dispersion. In our model, this trade-off is nonexistent because the planner uses both margins of demand as instruments to achieve both objectives.

Even though our model features a margin of demand that is not necessarily associated with market power, welfare losses are potentially large. This result follows from the observation that the endogenous allocation of customers pushes the Pareto frontier of our economy beyond what models without this extensive margin would suggest. While

[2] Since this model does not have the parameter that governs endogenous customer acquisition, it has one fewer parameter to calibrate. As a result, we drop the correlation between sales and non-production costs that our identification strategy relates to this parameter.

the uniform allocation of customers across firms is still feasible, the planner improves on this allocation by concentrating customers among more productive firms. Hence, welfare losses could be large if the equilibrium allocation of customers is sufficiently distorted. Thus, our analysis unveils a novel source of efficiency losses due to the *misallocation of customers*.

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

Literature review Our paper is closely connected to the literature that emphasizes the macroeconomic significance of the customer margin in firm growth and market share (Foster, Haltiwanger, and Syverson, 2015, Hottman, Redding, and Weinstein, 2016). Notably, recent research by Einav, Klenow, Levin, and Murciano-Goroff (2022) documents that about 80% of firms' sales variation arises from the customer margin, while Fitzgerald, Haller, and Yedid-Levi (2016), Argente, Fitzgerald, Moreira, and Priolo (2021) demonstrate that firms mainly grow through expansionary activities, rather than through reducing markups early in their life cycle. Our main contribution to this literature is to investigate the relationship between markups and different demand margins. Specifically, we document that once the market share is decomposed to the extensive and intensive margins, markups correlate only with firms' *average sales per customer*. Collectively, these facts paint a wholesome picture of firm growth: Firms mainly grow via the customer margin through non-price-related expansionary activities, with their markups tied only to their average sales per customer.

Based on these facts, our theoretical framework contributes to the literature on vari-

able markups by connecting models of firm growth through expansionary activities (e.g., Arkolakis, 2010) to models of endogenous markups at the intensive margin (e.g., Atkeson and Burstein, 2008).[3] Our contribution is to show that these two ingredients interact in a nontrivial way: Variable markups create differential incentives for firms to invest in their customer bases through non-price activities. These incentives lead to an equilibrium relationship between market shares and markups, but one that has different implications for the misallocation of resources relative to models that lack either ingredient. Finally, a notable recent work in this area is by Cavenaile, Celik, Perla, and Roldan-Blanco (2023), who provide microfoundations for the role of advertising in targeting different types of customers and showing how that leads to higher markups when advertising is more targeted.

Given our focus on misallocation, our paper is also related to the literature that analyzes the role of misallocation of production inputs across firms in affecting aggregate TFP (Restuccia and Rogerson, 2008, Hsieh and Klenow, 2009). In particular, our focus on the misallocation of customers *across* firms with variable markups relates our work to that of Bornstein and Peter (2022), who study misallocation of customers *within* firms, as well as Edmond, Midrigan, and Xu (2022), Peters (2020), who study the misallocation consequences of variable markups in settings without customer acquisition. We contribute to this literature by highlighting a new source of distortions in aggregate productivity that stems from the misallocation of customers across firms.

Layout Section 2 presents our empirical analysis. Section 3 describes the model. Section 4 discusses the model calibration. Section 5 quantifies the efficiency losses, and Section 6 concludes.

2 Motivating Facts

This section documents two new motivating facts using micro-level data that emphasize the importance of customer bases for firm dynamics and price-cost markups. Briefly, we document that price-cost markups are correlated only with average sales per customer and are unrelated to the size of firms' customer bases, even though these firms mainly grow through their customer bases.

[3] For growth through expansionary activities, see also Drozd and Nosal (2012), Kaplan and Zoch (2020), Einav, Klenow, Levin, and Murciano-Goroff (2022), Argente, Fitzgerald, Moreira, and Priolo (2021). There is also an extensive literature on growth through pricing activities; see, e.g, Phelps and Winter (1970), Rotemberg and Woodford (1999). For the most recent work in this area, see Bornstein (2021) and its review of that literature.

2.1. Data Description

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

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

where m_i denotes the number of firm i's customers, p_i its price, and q_i the average quantity purchased per customer.

We use the Nielsen Homescan Panel, which is one of the few sources of data that allows us to measure the number of each firm's customers.[4] The data contain approximately 4.5 million barcode-level product sales recorded from an average of 55,000 households per year in the United States. Nielsen samples households and provides in-home scanners so that households can record their purchases of products with barcodes. A barcode is a unique universal product code (UPC) allocated to each unique product and is used to scan and store its information. Each household is assigned a sample weight—or a projection factor—by Nielsen based on 10 demographic variables to make the sample nationally representative. Nielsen assigns a broad product-group label for each product, such as pet food and school supplies, and records information about the retailer a household visited to purchase products at a given time. According to Nielsen, the Homescan Panel covers approximately 30% of all household expenditures on goods in the consumer price index (CPI) basket. The data we use cover the period 2004-2016.

Next, we incorporate firm-level balance sheet information from Compustat to analyze firms' cost structures. With the caveat that this dataset covers only publicly listed firms, it constitutes the main source of panel data for firm-level analysis in the US and has been used in the recent literature on price-cost markups (see, for example, De Loecker, Eeckhout, and Unger, 2020, Traina, 2019). Throughout the analysis, we focus on two measures of a firm's costs. 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 includes "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

[4]The dataset was made available by the Kilts Marketing Data Center at the University of Chicago Booth School of Business.

to customers". It records costs attributable to the production of the goods sold by a firm, and its typical categories are the cost of labor and intermediate inputs used in production. On the other hand, SGA expenses include "commercial expenses of operation (such as expenses not directly related to product production) incurred in the regular course of business...". This includes the costs incurred to sell and deliver products and services and the costs to manage the company; typical categories are advertising, marketing, shipping, and research and development, among others.

Finally, we combine the Nielsen database with GS1 US Data Hub to group individual products according to their producing firms and to merge in firm-level information from Compustat. GS1 is the business entity that provides barcodes for products and records the firm name for each UPC available in the Nielsen data. The definition of a firm is based on the unit that purchased barcodes from GS1. Therefore, a firm in our data corresponds to either a manufacturer or a retailer. This merge procedure provides a unique link between customer- and producer-level data for each firm.[5] Although we only have 332 firms in the Nielsen-Compustat matched dataset, these cover approximately 25% of total sales in Nielsen. Section SM.1 and Table SM.1.1 in Supplemental Materials provide detailed descriptions of the data-cleaning procedure and the merged dataset.

2.2. The Relationship between Firm Size and Markups

We start by revisiting the predictions of a large class of models that relate firms' relative market power to their relative size. These models predict that larger firms charge higher markups (see Edmond, Midrigan, and Xu, 2022, Burstein, Carvalho, and Grassi, 2020, for supporting evidence). Since in most of these models every firm produces only one product, we can take this prediction to the data at either the firm level or the product level. We do both: We first present results at the firm level and then provide evidence that these results extend to the product level.

Evidence Based on Firm-level Markups Our decomposition of firms' sales in Equation (2.1) raises the following question: Which margin of sales captures the relationship between relative size and markups? To answer this question we first need to measure firm-level markups. We follow the methodology of De Loecker, Eeckhout, and Unger

[5]We match the Nielsen-GS1 database with the Compustat database, following a procedure similar to Argente, Lee, and Moreira (2018). We use the "reclink" STATA software command based on the 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 a 0.99 matching rate. We manually check the company name for every observation and drop inconsistent matches.

(2020), in which markups are measured by the inverse variable cost share of sales multiplied by the output elasticity of those variable inputs. Because this methodology does not require information on all variable costs, we follow De Loecker, Eeckhout, and Unger (2020) and use data on COGS from Compustat as a measure of variable costs. In addition, since we are interested in *relative* markups within industries at a given point in time, we absorb the output elasticity term with sector-year fixed effects.[6] Thus, our regression specification is given by:

$$\ln(\text{Sales}/\text{COGS})_{it} = \alpha_1 \ln p_{it} q_{it} + \alpha_2 \ln m_{it} + \lambda_{s,t} + \varepsilon_{it}, \tag{2.2}$$

where $(\text{Sales}/\text{COGS})_{it}$ is the Sales-to-COGS ratio of firm i at time t. The sector-year fixed effects $\lambda_{s,t}$ absorb all the variation at the sector-year level, which allows us to interpret the markup measure in *relative* terms and the size variables (average sales per customer and the number of customers) in terms of *market shares*.

Table 1: Markups, Sales per Customer, and Number of Customers

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

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

Our first motivating fact is that *firms' markups are highly correlated with their average sales per customer but are unrelated to their number of customers*, as reported in Table 1.

[6] The main challenge in the estimation of markups lies in estimating the output elasticity using only data on firms' revenues (see Bond, Hashemi, Kaplan, and Zoch, 2021). Given our focus on relative markups and our log regression specification, we do not need to estimate output elasticities. Instead, our regression specification incorporates a set of fixed effects that absorb these output elasticities. The underlying assumption we make, which is standard in the literature, is that the output elasticity with respect to COGS remains constant across firms within an industry and/or time.

Results are robust to including different combinations of year and sector fixed effects, which shows that the identified relationships hold within sectors by year.[7] With the inclusion of industry-time fixed effects, our results show that firms that charge higher markups have a higher market share in terms of average sales per customer; this is expected, given existing theories and previous evidence. However, the most novel result is that markups are not associated with firms' market shares in terms of the number of customers.[8] Thus, the relevant notion of relative size for markups is based on average sales per customer.[9]

Evidence Based on Product-level Markups Our results in Table 1 are limited by the scope of the Compustat dataset, which focuses on large public firms and firm-level markups. To demonstrate the external validity of these results, we perform two complementary analyses by using alternative measures of markups and sales. In the first approach, we measure the retailer-product-level markup as the difference between the retailer-product-level price available from the Nielsen Homescan Panel data and the wholesale cost obtained from the Nielsen PromoData, which is based on Gopinath, Gourinchas, Hsieh, and Li (2011) and Stroebel and Vavra (2019). In the second approach, we analyze the relationship between markups and each margin of demand by exploiting the rich dimensions of the data that allow us to control for marginal costs through an extensive set of fixed effects. This approach is similar to that of Fitzgerald, Haller, and Yedid-Levi (2016) and is valid under the assumption of common marginal costs across different subsets of observations. Both approaches include product fixed effects, which allow us to analyze the relationships within a given product. Appendix A.3 provides detailed descriptions of

[7] The sector fixed effects allow us to show that firms that sell more per customer have higher price markups *within* their sector. The advantage of this firm-level analysis is that it allows us to measure relative markups directly. However, it raises the question of whether this relationship is coming from within-firm products or is due to compositional effects across products. Since this approach measures price markups at the firm level in a limited sample, we cannot include fixed effects for products available in the Nielsen data or for more disaggregated SIC codes. However, as discussed below, we use two alternative approaches to measure markups at the product (UPC) level and include more disaggregated product fixed effects in Appendix A.3. We find that the reported relationships hold within products, which supports the view that they do not arise from compositional effects.

[8] One source of concern might be measurement error in a firm's customer base leading to attenuation bias. Two points alleviate this concern. First, the estimated coefficients are precisely estimated; i.e., we are finding a precise zero association. Second, in Appendix A.1, we use the lagged number of customers ($\ln m_{it-1}$) as an instrument for $\ln m_{it}$, which provides consistent estimates under classical measurement error. We find a strong first stage and the point estimates are similar to those in Table 1.

[9] Appendix A.2 presents results of an alternative specification that includes both total sales and average sales per customer as regressors. Given the decomposition in Equation (2.1), it is not surprising that markups are only associated with the average sales per customer but not with total sales.

both approaches.

Although we switch the focus to product-level markups and a broader sample of products and firms, both alternative approaches generate results consistent with those in Table 1: Markups are positively associated with average sales per customer, but not with the number of customers.

2.3. The Role of Customer Base for Sales Growth

Armed with the empirical evidence that shows how each margin of demand is associated with markups, we now document that the main source of variation in firm sales is the variation in the number of customers rather than average sales per customer. Table SM.1.2 in Supplemental Materials presents summary statistics of the customer-firm matched data. A quick glance of the Nielsen-GS1 data already reveals that much of the firm-product group-year sales are driven by the number of customers, not by average sales per customer: More than 500,000 customers spend only approximately $10 for each product group and firm per year on average.

To formalize this point, we follow Equation (2.1) and decompose the variance of log sales into the variances of log average sales per customer and log number of customers, as well as the covariance between these two components. Table 2 documents our second motivating fact: *Firms mainly grow by acquiring new customers instead of increasing their average sales per customer*. The number of customers accounts for approximately 80% of the variation in sales across firms. Average sales per customer, on the other hand, accounts for approximately 11% of the variance of sales, with the covariance accounting for the rest.[10] These results parallel the findings in contemporary work by Einav, Klenow, Levin, and Murciano-Goroff (2022). Using transaction-level data for a broad set of industries, they document that differences in customer bases account for 74% of sales variation across merchants. This shows that our finding extends beyond the consumer packaged goods sector and is representative of similar patterns in a wider set of industries.

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

Table 2: Decomposing the Variance of Sales

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

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

In addition to decomposing sales in the cross-section of firms, we find that the acquisition of new customers is also the main driver of firms' sales growth. As firms enter the economy, they can grow either by selling more per customer or by selling to more customers. To document this fact, we analyze firm growth patterns after entry. We mark a firm's entry as the time when it appears in our data for the first time. To be conservative, we drop entry events that occurred in the first 4 years in the dataset.[11] To quantify the importance of each margin, we estimate the following equation:

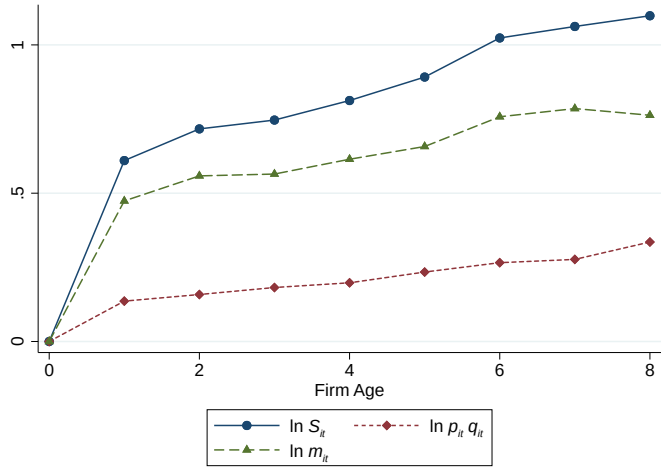
$$\ln S_{it} = \sum_{a=1}^{8} \delta_a \mathbf{1}(\text{age}_{it} = a) + \lambda_i + \lambda_t + \varepsilon_{it}, \tag{2.3}$$

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

Figure 1 plots the log sales as a function of firm age ($\hat{\delta}_a$ as a function of a) and decomposes it into the log number of customers and log average sales per customer. As a firm's sales grow over time, both margins of demand also increase. However, regardless of the firm's age, sales growth is mostly attributed to the increase in the number of customers. At age 1, differences in the number of customers explain approximately 78% of differences in sales, whereas average sales per customer explain approximately 22% of sales. Although the importance of the number of customers decreases as firms become older, on average, it still accounts for approximately 70% of sales for the maximum firm age observed in the data. Results are robust to including only those firms that survive at least 3 or 5 consecutive years and analyzing average monthly sales as the dependent variable, which accounts for staggered entry throughout the year. Since the degree of

[11] 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 drop the years 2004-2007 to render the analysis conservative. Thus, the maximum firm age in our sample of entrants is 8 years.

Figure 1: Decomposition of Firm Sales Growth by Firm Age



Notes: This figure plots the average firm sales, sales per customer, and the number of customers for each firm-age based on Equation (2.3), after controlling for firm and year fixed effects. The blue circled line shows the results for log sales, the red diamond line for the log average sales per customer, and the green triangle line for the log number of customers. There are 40,442 observations and 9,990 firms that newly enter the economy starting from the year 2008 in the Nielsen-GS1 data. All estimates are normalized based on age 0. All variables are projection-factor adjusted.

durability of a product might affect the ability of firms to grow through different margins, we repeat the analysis by splitting products according to their durability and find similar patterns within both subsamples. See Appendix A.4 for further details.

The fact that firms mainly grow through the extensive margin, which is not associated with their markups, begs the question: To what extent do firms control their growth through different sales margins? To investigate this, in Appendix A.5 we use data on SGA expenses, which in part capture firms' expenditures on expansionary activities, and find that (1) firms that spend more non-production costs have higher sales, and (2) these costs are associated with the number of new customers firms acquire, but not with the number of customers they retain or average sales per customer. In summary, the evidence shows that firms' SGA expenses contribute to relative firm size only through customer acquisition.

All these facts together suggest that by spending on expansionary activities, firms grow through a margin of demand that *does not* correlate with markups—a distinction not previously studied in the macroeconomics literature. In the next section, we develop a model that can account for these facts and explore its implications.

3 Model

In this section, we present a model with variable markups and endogenous customer acquisition that is consistent with our motivational facts in Section 2. We then use this model to quantitatively investigate the implications of endogenous customer acquisition for markup dispersion, misallocation, and welfare.

3.1. Setup

Time is discrete and is indexed by $t \in \{0,1,2,\ldots\}$. There is a representative household with a continuum of individual members denoted by $j \in [0,1]$. A continuum of firms, indexed by $i \in N_t$, produce weakly substitutable goods in a representative industry. With slight abuse of notation, we use N_t to denote both the set and the measure of these firms.

3.1.1. Households. The representative household supplies labor in a competitive labor market and demands different varieties produced by firms at given prices. We let $m_{i,t}$ denote both the measure and the set of a variety i's customers and write $j \in m_{i,t}$ when member j is a customer of i. Household members jointly maximize their utility when the utility of their consumption baskets is aggregated by a *Kimball aggregator*, subject to their budget constraint:

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

$$s.t. \int_0^{N_t}\int_0^1 \mathbf{1}_{\{j\in m_{i,t}\}} \Upsilon\left(\frac{c_{i,j,t}}{C_t}\right) dj di = 1 \tag{3.2}$$

$$\int_0^{N_t}\int_0^1 p_{i,t} c_{i,j,t} dj di \leq W_t L_t + \int_0^{N_t} \Pi_{i,t} di - T_t. \tag{3.3}$$

Here $c_{i,j,t}$ is the consumption of member j from variety i; $p_{i,t}$ is the price of variety i; C_t is aggregate consumption; L_t is the total labor supply; W_t is the wage; $\Pi_{i,t}$ is the profit of i; and T_t is a lump-sum tax. The function $\Upsilon(.)$ is strictly increasing and concave with $\Upsilon(1) = 1$.[12] It follows that all customers of i choose to purchase the same amount implied by the following demand function:

$$\frac{c_{i,j,t}}{C_t} = q_{i,t} \equiv \Upsilon'^{-1}\left(\frac{p_{i,t}}{P_t D_t}\right) \mathbf{1}_{\{j\in m_{i,t}\}}. \tag{3.4}$$

Here $q_{i,t}$ denotes the relative demand per matched customer of variety i. Moreover,

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

[12] 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 the price of the aggregate consumption good, which, henceforth, we normalize to one.[13] The homogeneity of $q_{i,t}$ across all customers of firms follows from the homogeneity of preferences. In Section SM.2 in Supplemental Materials, we introduce an extension of the model with heterogeneity in tastes and provide motivational evidence for why we abstract from them. Therefore, the household's total demand for variety i is *proportional* to the number of its customers, and given by the demand function:

$$\begin{aligned} c_{i,t} &\equiv \int_0^1 \mathbf{1}_{\{j \in m_{i,t}\}} c_{i,j,t}\, dj = m_{i,t} q_{i,t} C_t \\ \implies \ln(S_{i,t}) &\equiv \ln(p_{i,t} c_{i,t}) = \ln(m_{i,t}) + \ln(p_{i,t} q_{i,t}) + \ln C_t \end{aligned} \tag{3.5}$$

where $c_{i,t}$ is total demand for variety i, and $q_{i,t}$ is the relative demand per customer in Equation (3.4). Also, the expression on the right shows how the model delivers the same decomposition of sales to the number of customers and demand per customer as in our empirical analysis in Equation (2.1).[14]

For the functional form of $\Upsilon(.)$, we use the Kimball aggregator of Klenow and Willis (2016):

$$\Upsilon(q) = 1 + (\sigma - 1) e^{\frac{1}{\eta}} \eta^{\frac{\sigma}{\eta} - 1} \left[\Gamma\left(\frac{\sigma}{\eta}, \frac{1}{\eta}\right) - \Gamma\left(\frac{\sigma}{\eta}, \frac{q^{\frac{\eta}{\sigma}}}{\eta}\right) \right], \tag{3.6}$$

where $\sigma > 1$ and $\eta > 0$ control the demand elasticities and super-elasticities, as we discuss below, and $\Gamma(.,.)$ is the incomplete Gamma function.[15] This specification for $\Upsilon(.)$ is a generalization of a CES aggregator with substitution elasticity σ, which is nested when $\eta = 0$. Using this functional form in Equation (3.4), we obtain the following relative demand per customer for firm i at time t:

$$q_{i,t} = \left[1 - \eta \ln\left(\frac{p_{i,t}}{D_t (1 - \sigma^{-1})} \right) \right]^{\frac{\sigma}{\eta}}. \tag{3.7}$$

Intuitively, this demand function is a smoothed version of a kinked demand curve (Dotsey and King, 2005, Basu, 2005), in which a customer's demand is more price sensitive at

[13] 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, we could characterize the equations that pin down P_t and D_t in terms of prices. The equations that determine P_t and D_t are:

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

[14] In the empirical analysis, the $\ln C_t$ term would be absorbed by the industry fixed effects.

[15] The incomplete Gamma function is given by $\Gamma(s,x) \equiv \int_x^\infty t^{s-1} e^{-t}\, dt$.

higher relative prices—i.e., demand satisfies Marhalls's second law of demand. As we show below, firms with larger demand per customer face lower elasticities and charge higher markups. To observe this, let us consider the demand elasticity $\varepsilon_{i,t}$, and super-elasticity, $\varepsilon^{\varepsilon}_{i,t}$:

$$\varepsilon_{i,t} \equiv -\frac{\partial \ln(c_{i,t})}{\partial \ln(p_{i,t})} = \sigma q_{i,t}^{-\frac{\eta}{\sigma}}, \quad \varepsilon^{\varepsilon}_{i,t} \equiv \frac{\partial \ln(\varepsilon(q_{i,t}))}{\partial \ln(p_{i,t})} = -\frac{\eta}{\sigma}\varepsilon_{i,t} \leq 0, \tag{3.8}$$

where we see that demand elasticity $\varepsilon_{i,t}$ is a decreasing function of the relative demand per customer, formally shown by the negative sign of the super-elasticity $\varepsilon^{\varepsilon}_{i,t}$ as long as $\eta > 0$.

Finally, to conclude households' optimality conditions, the household's labor supply is characterized by the following standard intratemporal Euler equation: $\xi L_t^{\psi} = W_t C_t^{-\gamma}$.

3.1.2. Dynamics of Customer Bases. We model the dynamics of customer bases following our empirical findings in Section 2. In particular, as we document in Appendix A.5, firms' SGA expenses correlate with the acquisition of new customers (but not the retention of old customers).

Motivated by this evidence, we assume that firms can engage in expansionary activities, such as advertising campaigns or increasing the availability of their goods, to attract *new customers*. In addition, two processes in the model separate customers from firms: at the end of each period, (1) all customers of exiting firms separate, and (2) customers of incumbent firms separate at an exogenous rate of $\delta \in [0,1]$. We assume that the total mass of matches is fixed over time and, without loss of generality, normalize this mass to one. This implies that while expansionary activities affect the distribution of customers across firms, it does not increase the total number of customers who buy from an industry (as in Einav, Klenow, Levin, and Murciano-Goroff, 2022). We view this as a conservative benchmark, as we discuss in Section 5.

As for the dynamics of new matches, we assume that operating firm i at time t posts $a_{i,t} \geq 0$ ads to acquire new customers. Every unmatched member then draws an ad from the pool of all available ads and is matched to the firm they draw. Therefore, the number of new customers firm i acquires at time t is proportional to the number of ads it posted relative to the total number of ads posted by all firms.[16] Hence, firm i's customer base

[16] Note that we have modeled customer acquisition through expansionary activities that operate independent of firms' pricing decisions, and thus abstract away from customer acquisition through lower markups/prices. This is motivated by the evidence in Fitzgerald, Haller, and Yedid-Levi (2016), who conclude that firms do not manipulate prices to shift demand by documenting that after entering a new market, their markups remain the same while their quantities grow. Our findings in Table 1 and Appendix A.5 also

evolves according to

$$m_{i,t} \leq (1-\delta) m_{i,t-1} + \frac{a_{i,t}}{P_{m,t}}, \tag{3.9}$$

where the inequality captures the notion that there is free disposal of customers, should the firm choose to exercise that option. Moreover, $P_{m,t}$ is the endogenous conversion rate of ads to customers. This is the number of ads needed for a firm to get one new customer, determined in the equilibrium to clear the matching market:

$$\int_{i \in N_t} m_{i,t} di = 1 \implies P_{m,t} = \frac{\int_{i \in N_t} a_{i,t} di}{1 - (1-\delta) \int_0^{N_t} m_{i,t-1} di}. \tag{3.10}$$

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

3.1.3. Firms. On the firm side, we assume endogenous entry and exit with an order of events as summarized in Figure 2. We provide a detailed description of these decisions below.

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:

$$\ln(z_{i,t}) \sim \mathcal{N}(\bar{z}_{ent}, \sigma_z^2). \tag{3.11}$$

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

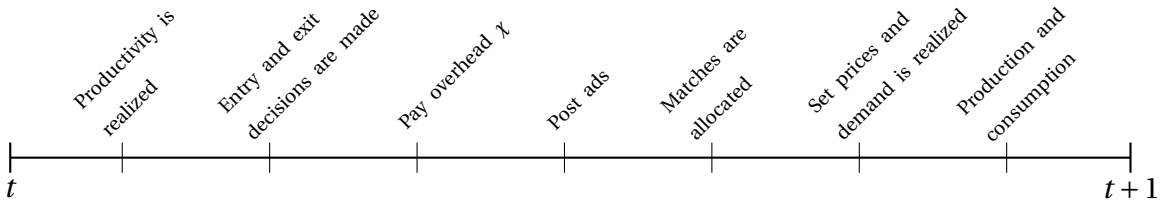
$$\ln(z_{i,t}) = \rho \ln(z_{i,t-1}) + \epsilon_{i,t}, \quad \epsilon_{i,t} \sim \mathcal{N}(0, \sigma_z^2). \tag{3.12}$$

With new productivities drawn, each incumbent or potential entrant then decides whether to stay in the economy or to drop out.[17] We refer to this decision by $\mathbf{1}_{i,t} \in \{0,1\}$, with 1 being an indicator for entering or staying. Finally, all the incumbent firms who decided to stay draw i.i.d. Bernoulli survival shocks, $\nu_{i,t}$, that are equal to 1 with probability $\nu \in [0,1]$, and drop out if $\nu_{i,t} = 0$.

support the importance of customer acquisition through expansionary activities by showing that markups are not correlated with the size of firms' customer bases conditional on sales per customer, but expenses related to expansionary activities are.

[17] Following Clementi and Palazzo (2016) and Ottonello and Winberry (2020), we allow the mean of incumbents' productivity distribution—normalized to 0—to differ from 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 2: Order of Events



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

Expansionary Activities, Pricing, and Production Firms that stay or enter the economy pay an overhead cost of $\chi > 0$ in units of labor at each period. Also, firms use labor to produce ads using the technology $a_{i,t} = l_{i,s,t}^{\phi} \geq 0$, where $l_{i,s,t}$ denotes the amount of labor allocated to advertising activities. The firm's customer base then evolves according to Equation (3.9), where $m_{i,t-1} \equiv 0$ for firms that entered at time t. Moreover, $\phi \in [0,1]$ is the degree of decreasing returns to advertising. Once new customers are acquired, firms' demands are realized as in Equation (3.5). Taking this demand as given, firms then choose the prices. Each firm i then produces to meet its realized demand using the production function $y_{i,t} = z_{i,t} l_{i,p,t}^{\alpha}$, where $z_{i,t}$ is the firm's productivity, and $l_{i,p,t}$ is its labor demand for production. Finally, $\alpha \in [0,1]$ is the degree of decreasing returns to production.

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

$$v_t(m_{i,t-1}, z_{i,t}) \equiv \max_{\substack{(q_{i,\tau}, l_{i,s,\tau}, \\ l_{i,p,\tau}, \mathbf{1}_{i,\tau})_{\tau=t}^{\infty}}} \mathbb{E}_t \sum_{\tau=t}^{\infty} (\beta\nu)^{\tau-t} \left(\prod_{h=t}^{\tau} \mathbf{1}_{i,h} \right) \left(\frac{C_\tau}{C_t} \right)^{-\gamma} \left[\overbrace{D_\tau \Upsilon'(q_{i,\tau}) y_{i,\tau}}^{\text{price}} - \underbrace{W_\tau l_{i,p,\tau}}_{\text{COGS}} - \underbrace{W_\tau (l_{i,s,\tau} + \chi)}_{\text{SGA expenses}} \right] \tag{3.13}$$

$$\text{subject to} \quad y_{i,\tau} = m_{i,\tau} q_{i,\tau} C_\tau = z_{i,\tau} l_{i,p,\tau}^{\alpha} \tag{3.14}$$

$$m_{i,\tau} \leq (1-\delta) m_{i,\tau-1} + \frac{l_{i,s,\tau}^{\phi}}{P_{m,\tau}}, \quad l_{i,s,t} \geq 0. \tag{3.15}$$

The problem states that firm i maximizes the expected discounted stream of its profits subject to its demand curve in Equation (3.14), the law of motion for customers in Equation (3.15), and the nonnegativity of labor allocated to advertising. We have also labeled the terms in the profit function as *total sales*, *COGS*, and *SGA expenses*, which we use later to map the model to the data.

3.2. Characterization of Firms' Decisions

In this section, we characterize the firms' optimal decision rules for pricing, expansionary activities, and entry and exit. All proofs are in Appendix B.

Prices and Markups Conditional on decisions for entry/exit and customer acquisition, firms' pricing decisions have a static nature. Formally, firm i's optimal price at t is:

$$p_{i,t} = \underbrace{\frac{\varepsilon_{i,t}}{\varepsilon_{i,t}-1}}_{\text{markup}} \times \underbrace{\alpha^{-1}\frac{W_t l_{i,p,t}}{y_{i,t}}}_{\text{marginal cost}}. \tag{3.16}$$

This expression shows that despite the presence of variable customer acquisition costs, the usual relationship between the labor share and the markup also holds in this model. This verifies our use of De Loecker, Eeckhout, and Unger (2020) methodology to identify markups from the Compustat data. Moreover, it is also important to note that the firm's elasticity of demand, $\varepsilon_{i,t}$, is itself a function of demand per customer in Equation (3.7) and varies with the firm's pricing choice. Therefore, as long as η is not zero, the optimal markup of the firm varies with its marginal cost, which leads to the following lemma.

Lemma 1. At a given time t, firms with higher marginal costs charge higher prices and lower markups. Formally, let $\mu_{i,t}$ denote a firm's markup and $mc_{i,t}$ its marginal cost. Then, the elasticities of markups and prices to marginal costs are:

$$\frac{\partial \ln(p_{i,t})}{\partial \ln(mc_{i,t})} = \frac{1}{1+\eta\sigma^{-1}\varepsilon_{i,t}(\mu_{i,t}-1)} \geq 0 \tag{3.17}$$

$$\frac{\partial \ln(\mu_{i,t})}{\partial \ln(mc_{i,t})} = -\frac{\eta\sigma^{-1}\varepsilon_{i,t}(\mu_{i,t}-1)}{1+\eta\sigma^{-1}\varepsilon_{i,t}(\mu_{i,t}-1)} \leq 0. \tag{3.18}$$

Equation (3.17), which is also known as the *incomplete pass-through* property of Kimball demand (see, e.g., Gopinath and Itskhoki, 2010, Amiti, Itskhoki, and Konings, 2019), shows that a 1% increase in the marginal cost of a firm increases their price by less than 1%. The intuition is that firms with higher marginal costs need to charge higher relative prices to keep their positive margin. But at such new prices, demand is more elastic and optimal markups are lower, in contrast to a CES demand in which demand elasticities are constant and pass-through is complete. With these variable markups, we obtain the following proposition.

Proposition 1. At any t, firms with *higher relative sales per customer* charge *higher markups*:

$$\frac{d\ln(\mu_{i,t})}{d\ln(p_{i,t}q_{i,t})}\bigg|_t = \eta\sigma^{-1}\mu_{i,t}(\mu_{i,t}-1) \geq 0 \tag{3.19}$$

Proposition 1 shows that firms with higher sales per customer charge higher markups, which links the model to our empirical fact on this relationship in Table 1. Intuitively, firms with higher sales per customer must have lower relative prices and thus can charge higher markups because demand is less elastic at such prices. Note, however, that firms with higher sales per customer do not necessarily have to be larger in terms of total sales, since the size in our model also depends on the number of firms' customers. Due to the dynamic evolution of the customer base, for any given market share there is a distribution of firms with different sales per customer holding that market share. As a result, fixing market share, firms with more customers must also sell less per customer, and as a corollary of Proposition 1, charge lower markups.

Corollary 1. Conditional on relative total sales, firms with more customers charge lower markups:

$$\frac{d\ln(\mu_{i,t})}{d\ln(m_{i,t})}\Big|_{\frac{p_{i,t}y_{i,t}}{\int_{i\in N_t} p_{i,t}y_{i,t}di},t} = -\eta\sigma^{-1}\mu_{i,t}(\mu_{i,t}-1) \leq 0 \tag{3.20}$$

Corollary 1 highlights the main departure of our paper from the literature on variable markups, in which customer acquisition is not modeled explicitly and customers are *homogeneously* distributed across firms. In such models, the fact that all firms are on the same demand curve implies that relative size (market share) is a sufficient statistic for firms' markups. However, in our model, firms can hold the same relative size either because they sell more per customer or because they have more customers, which implies that relative size is no longer a sufficient statistic for market power.

But what determines the *unconditional* relationship between relative size and markups in this model? To answer this question, we need to characterize firms' optimal expansionary activities.

Optimal Expansionary Activities A key feature of our model is that firms internalize the decision to acquire customers. For this decision, while the marginal cost of a new customer is determined by the amount of labor the firm needs to employ to find it, its benefit is closely linked to the firm's market power and the amount of additional demand from that customer. The following proposition formulates the optimality condition for firms' advertising decisions in terms of this cost-benefit analysis.

Proposition 2. The optimal customer acquisition strategy of a firm is characterized by

$$\underbrace{\phi^{-1}\frac{W_t l_{i,s,t}}{m_{i,t}-(1-\delta)m_{i,t-1}}}_{\text{marginal cost of a new customer}} = \mathbb{E}_t \sum_{\tau=t}^{\infty} \underbrace{\left[(\nu(1-\delta))^{\tau-t}\prod_{h=t}^{\tau}\mathbf{1}_{i,\tau}\right]}_{\text{probability of match survival}} \times \underbrace{\beta^{\tau-t}\left(\frac{C_\tau}{C_t}\right)^{-\gamma}\left(\mu_{i,\tau}-1\right)mc_{i,\tau}q_{i,\tau}C_\tau}_{\text{discounted (gross) marginal profit per customer}}. \tag{3.21}$$

Equation (3.21) shows that firms optimally equate the cost of acquiring the marginal customer to the net present value of the gross profits earned from them for the duration of the match. Since the marginal profits generated by a new customer are increasing in the firm's markups, 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 *but endogenous* relationship between markups and the size of firms' customer bases in the equilibrium. This is in contrast to a model with an exogenous customer base or a representative household in which any relationship between markups and relative size is dictated by the shape of the exogenous demand curves. The endogenous nature of this relationship in a model with customer acquisition hints at its different counterfactual implications, which we will discuss in later sections.

Entry and Exit Policies A potential entrant enters and an incumbent stays if their value, specified in Equation (3.13), is positive: $v_t(m_{i,t-1}, z_{i,t}) > 0$. It follows that for any m_{-1}, there is a threshold $z^*(m_{-1})$ such that firms with higher productivity than $z^*(m_{-1})$ stay or enter (see Hopenhayn, 1992).

3.3. Equilibrium

An equilibrium is defined as (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 $\{(p_{i,t}, y_{i,t}, m_{i,t}, l_{i,p,t}, l_{i,s,t})_{i\in N_t}\}_{t\geq 0}$; and (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 (3.13),
3. labor and matching markets clear: $L_t = \int_{i\in N_t}(l_{i,p,t} + l_{i,s,t} + \chi)di, \quad 1 = \int_{i\in N_t} m_{i,t}di$,
4. the set of operating firms, N_t, evolves according to

$$N_t = \{i \in \Lambda_t \cup N_{t-1} : \mathbf{1}_{i,t}\nu_{i,t} = 1\},\ N_{-1} \text{ given.} \tag{3.22}$$

Solution Method We solve the model globally by combining collocation methods and nonstochastic simulation to approximate the distribution of firms over $(z_{i,t}, m_{i,t-1})$. Section SM.3 in Supplemental Materials describes the recursive formulation of the firm's problem and the computational algorithm that finds the steady state of this economy.

4 Quantitative Analysis

To quantify the implications of customer allocation with variable markups, we calibrate the steady state of the model using the Simulated Method of Moments and matching several micro and macro moments related to firm dynamics in the US economy.

4.1. Calibration Strategy

To provide an overview of our calibration strategy, the new and most relevant parameter to calibrate is ϕ, the returns to scale in advertising—which influences the relationship between a firm's relative size and market power by determining the size of its customer base.[18] As we discuss below, this parameter is identified by the relationship between sales and firms' expansionary activities included within the SGA expenses. The fact that the latter is only available from Compustat poses a challenge because it contains only a subset of the firms in the economy. To address this challenge, 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 based on data from Compustat, we apply a filter to the model-simulated data to account for the selection into Compustat based on firm size and age.[19]

Fixed Parameters We set the length of a period to 1 year. Panel A of Table 3 presents the set of parameters that are externally fixed. We set the subjective discount factor β to 0.96. The elasticity of intertemporal substitution γ is set to 2. We set the inverse of the Frisch elasticity of labor supply to $\psi = 1$ and the labor coefficient in the production function to $\alpha = 0.64$. In the calibration exercise, we normalize the measure of potential entrants λ

[18] Note that our model does not require firms to spend on customer acquisition and grow through the extensive margin of demand, and nests the conventional model with exogenous customer bases as a special case. When $\phi \to 0$, every firm receives the same flow of customers in each period, without having to spend on $l_{i,s,t}$. If, in addition, $\delta = 1$, then all firms have the same stock of customers in every period.

[19] That is, to compute equivalent moments, we restrict the simulated sample of firms to those that are at least 7 years old, as in Ottonello and Winberry (2020), 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). 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.

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 of customers in the Nielsen-GS1 matched dataset in 2012.[20] This value of the repurchasing probability is similar in other industries based on evidence from the marketing literature.[21]

Calibrated Parameters We jointly calibrate the remaining 8 parameters by the simulated method of moments (SMM).[22] These parameters can be grouped in three sets: those shaping firms' cost structure (ϕ and χ), their demand (σ, and η), and their life cycle and shock structure ($\rho_z, \sigma_z, \bar{z}_{ent}$ and ν). Although these parameters are jointly identified by all moments, we provide below a discussion of which moment should intuitively be more relevant to identify each parameter. We formalize this discussion in Section SM.4.1 in Supplemental Materials 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. The model counterpart of this ratio for firm i is $\frac{W_t l_{i,p,t}}{W_t l_{i,p,t} + W_t (l_{i,s,t} + \chi)} \equiv \frac{COGS_{i,t}}{COGS_{i,t} + SGA_{i,t}}$. Intuitively, a larger fixed cost χ, ceteris paribus, should increase a firm's total costs and drive down this ratio. To measure this ratio, we use data from Compustat.

To identify the elasticity ϕ, we exploit the observed relationship between SGA and sales in Compustat. Proposition 3 illustrates the source of identification in the special case with $\delta = 1$.

[20] More specifically, define $\text{Sales}_{i,g,t}$ as the total expenditure of (projection-factor adjusted) households that purchase products made by firm i in group g at time t. Define the probability of repurchasing firm's products as $s_{i,g,t} = \frac{\text{Sales}_{i,g,t-1,t}}{\text{Sales}_{i,g,t-1}}$, where $\text{Sales}_{i,g,t-1,t}$ 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, in which the weights are the expenditure in firm-group bins across all years.

[21] For example, the repurchase probability is 0.7 in the automotive industry based on survey data used by 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 services industry. Finally, Bornstein (2021) estimates an annual retention probability of 0.85 for the top two largest firms in each product category from the Nielsen data. If we also restrict the sample to the top two firms, our retention measure increases to 0.84.

[22] More specifically, we calibrate the model by choosing a set of parameters $\mathscr{P} = (\phi, \chi, \sigma, \eta, \rho_z, \sigma_z, \bar{z}_{ent}, \nu)$ that minimizes the SMM objective function $\left(\frac{\boldsymbol{m}_m(\mathscr{P})}{\boldsymbol{m}_d} - 1\right)' \boldsymbol{W} \left(\frac{\boldsymbol{m}_m(\mathscr{P})}{\boldsymbol{m}_d} - 1\right)$, 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. Section SM.3.2 in Supplemental Materials provides the computational details of the calibration exercise.

Proposition 3. 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:

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

Equation (4.1) is obtained from the firm's optimality condition Equation (3.21) regarding customer acquisition, which shows that firms' expenditures on customer acquisition are increasing in their markups because they directly determine the returns from customer acquisition. Since ϕ is the returns to scale in acquiring new customers, it naturally follows that it should be identified from the sensitivity of SGA expenditures to firms' markups, which in the model are proportional to firms' Sales-to-COGS ratios. Proposition 3 formalizes this intuition in the special case of $\delta = 1$, and in Section SM.4 in Supplemental Materials we find a high sensitivity of ϕ to the same relationship; this indicates that our intuition also carries on for the general case with $\delta < 1$. Thus, we identify the sensitivity of SGA expenses to markups by regressing firms' SGA expenses on their sales while controlling for COGS and time fixed effects (which capture the fixed components of SGA expenses). We calibrate ϕ so that the regression coefficient on $Sales_{i,t}$ in the model-simulated data matches the regression coefficient obtained from Compustat.

To calibrate the parameters that shape firms' demand, we set the elasticity of substitution σ to match a COGS-weighted average markup of 1.25 computed from Compustat as in Edmond, Midrigan, and Xu (2022). While the level of markups identifies σ, the sensitivity of markups to relative size identifies η. Since we do not directly observe markups, following Edmond, Midrigan, and Xu (2022), we pin down the super-elasticity of demand η 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 directly proportional to the production markup $\mu_{i,t}$, which validates this approach. The following proposition confirms that a similar relationship holds in our model (see Section SM.4 in Supplemental Materials for the numerical mapping between η and this relationship).

Proposition 4. Suppose $\delta = 1$. Then, a firm's average revenue productivity of labor is given by

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

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

When $\eta = 0$, markups and the revenue productivity of labor are constant and indepen-

dent of sales. However, markups and revenue productivity become positively correlated with sales when $\eta > 0$ because the demand elasticity decreases with size in proportion to η. Therefore, the strength of the relationship between labor productivity and sales is informative about η, holding the other parameters fixed. We summarize this relationship with the regression coefficient of a sales-weighted OLS regression of relative revenue productivity of labor on relative sales of 0.036 for firms with relative sales greater than 1, as computed by Edmond, Midrigan, and Xu (2022) using aggregate data from SUSB.[23]

Finally, the parameters of the AR(1) productivity process for incumbent firms, σ_z and ρ_z, are set to match a standard deviation of annual employment growth of 0.415 from Elsby and Michaels (2013) and the unweighted distribution of within-industry relative sales from Edmond, Midrigan, and Xu (2022). 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).

Calibration Results The set of calibrated parameters is shown in Panel B of Table 3. 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. The value for the elasticity is standard and the value for the super-elasticity is close to estimates found in the literature (see Nakamura and Zerom, 2010). Finally, note that the calibrated value for the elasticity of the matching function $\phi = 0.533$ is close to a model-generated regression coefficient of 0.474. This similarity lends support to the identification argument provided in Proposition 3. Section SM.4.2 in Supplemental Materials shows that our calibrated model matches the targeted moments reasonably well.

[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 (e.g., production and advertising labor). The decision not to include χ is due to the fact that part of overhead costs are, in reality, not associated with labor costs (e.g., rent) and thus not included in the wage bill reported by SUSB. Ideally, we would use data on the subcomponents of SGA expenses in Compustat to compute the share of labor costs within SGA expenses. Unfortunately, a full disaggregation of SGA expenses is not available. To alleviate concerns about this choice, note that we target a moment based on a sample of relatively large firms (those with relative sales greater than 1), for which arguably the fixed overhead cost represents a smaller fraction of total costs.

Table 3: Model Parameters

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

Notes: This table shows the calibration of the model. Panel A contains parameters externally chosen. Panel B contains parameters internally calibrated to match moments presented in Table SM.4.1 and Figure SM.4.2 in Supplemental Materials.

4.2. Model Validation

We also provide overidentifying tests of the calibrated model regarding its ability to match relevant untargeted moments. We have previously documented that, in the data, the major source of cross-firm differences in sales is the size of their customer bases. This is verified in Table 4, which compares the variance decomposition of log sales in the model with the decomposition from the data. In the data, differences in the log of average sales per customer account for 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 driven by only one source of heterogeneity (i.e., the productivity shocks), the model naturally overpredicts the size of

the covariance term.[24]

Table 4: Sources of Sales Dispersion across Firms

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

Notes: This table provides a variance decomposition of firms' log sales. The first column reports the variance of the log sales per customer, $var\left(\ln p_{i,t}\Upsilon'(p_{i,t}/D_t)\right)$, relative to the overall variance of log sales. The second column reports the relative variance of the log number of customers, $var\left(\ln m_{i,t}\right)$. The last column reports the covariance between both terms, $cov\left(\ln m_{i,t}, \ln p_{i,t}\Upsilon'(p_{i,t}/D_t)\right)$. 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 strongly associated with market power (see Table 1). Next, we show that our model quantitatively reproduces this untargeted fact by regressing firms' markups on sales per customer and the size of their customer bases using model-simulated data:

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

Table 5 presents the results. The data show a significant relationship between markups and sales per customer and an economically as well as statistically insignificant relationship between markups and the size of the customer base. The model matches these facts fairly well and predicts that 1% higher average sales per customer are associated with 0.11% higher markups. This point estimate is between the baseline estimate of 0.06 reported in Table 1 and the estimate of 0.187 reported in the additional analysis in Appendix A.3. On the other hand, a 1% increase in the size of the customer base increases markups by only 0.02%.[25] Therefore, the model captures the differential roles of the intensive and extensive margins of demand on firms' markups documented in the data.

Finally, the model is able to generate firm dynamics similar to those observed in the data. Figure SM.4.6 in Supplemental Materials shows two model-based moments that were not explicitly targeted in the calibration exercise: a decreasing average exit rate

[24] The difference between the model and the data could be explained, for instance, by (unmodeled) orthogonal preference shocks that affect the size of the customer base.

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

Table 5: Sources of Dispersion in Sales and Markups

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

Notes: This table reports the results of an OLS regression of a firm's log markup ($\ln(\mu_{i,t})$) on log sales per customer ($\ln p_{i,t} q_{i,t}$) and log size of the customer base ($\ln m_{i,t}$). Column (1) reproduces the empirical estimates from Table 1. 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 or Year FE because we model a single "representative" industry in steady state.

by age and a decreasing average employment growth by age (as in the data; see, e.g., Haltiwanger, Jarmin, and Miranda, 2013). Intuitively, this is explained by the fact that entrants enter the economy with lower average productivity and no customer base. This makes young firms more likely to exit when faced with negative productivity shocks due to overhead costs and, conditional on staying, to grow more rapidly than older firms due to frontloaded focus on customer acquisition.

4.3. The Role of Endogenous Customer Acquisition

In this section, we investigate how endogenous customer acquisition changes the implications of the relationship between the size distribution of firms and their markups. To do so, we compare our calibrated model to an alternative model without customer heterogeneity that is calibrated to match the same data moments.

This alternative model—labeled "homogeneous customers" hereafter—corresponds to a version of our model in which all firms are exogenously matched with the representative household (i.e., $m_{i,t} = 1, \forall i$) and it has the same demand structure as in Klenow and Willis (2016), Edmond, Midrigan, and Xu (2022).[26] The results of this calibration, as well as results for its goodness of fit, are reported in Section SM.5 in Supplemental Materials.

[26] Since this model does not feature endogenous customer acquisition, it lacks the parameter ϕ that determines the returns to advertising. For this reason, in our recalibration exercise we drop the moment on the relationship between SGA expenses and sales, which was used to pin down ϕ.

Here, we discuss the main implication of this exercise. In Section SM.4.3 in Supplemental Materials, we also provide comparative statics results—i.e., comparisons under the same parameter values—between these two models to illustrate the mechanisms at play.

How does shutting down endogenous customer acquisition distort our interpretation of the data through the lens of the model? As shown in Figure 3, the recalibrated homogeneous customers model generates much lower markup dispersion relative to our baseline model while matching the same level of cost-weighted markup.[27]

The intuition behind this difference between the two models is closely related to how the two models match the distribution of sales across firms under demand curves that allow for heterogeneous markups. In this class of models, markup variation stems from the fact that at the intensive margin, a consumer's demand is less elastic at lower relative prices. Therefore, the maximum variation in markups these models can generate depends on how much variation there is in this elasticity across the demand curve, which is governed by the ratio η/σ as illustrated in Equation (3.8). Higher values of η/σ allow for higher variation in markups across firms, and both models nest CES demand (i.e., no variation in markups) when $\eta = 0$. Note, however, that a higher value of η/σ means that the demand elasticity of each customer declines more with relative prices, which implies that their quantity demanded increases by less as prices decline. Visually, a higher η/σ bends the right tail of the demand curve downward. In fact, it is well known that this property leads to a "choke quantity" at the intensive margin—i.e., a maximum quantity that is bought as relative prices approach zero (Edmond, Midrigan, and Xu, 2022). It is only when $\eta \to 0$ that this quantity goes to infinity, and we obtain the CES demand. By drawing this link, we would like to emphasize that the existence of a choke quantity is closely related to the amount of variation a model implies for markups across firms.

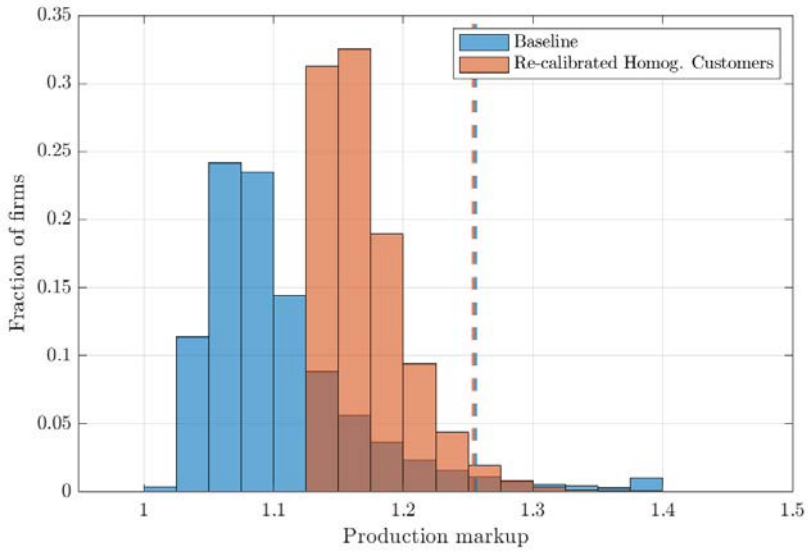
These choke quantities are exactly the root cause of why the homogenous customers model implies a much lower markup dispersion than our baseline model. To see why, note that *all* demand in the Homogenous customers model comes from the intensive margin; i.e., this model locates all the firms in the economy on the same demand curve. So, any variation in the size of firms in this model comes from these firms being at different parts of this single demand curve. Thus, a choke quantity puts an upper bound on how much variation such a model can generate in the size distribution of firms. As a result, the Homogenous customers model faces a fundamental trade-off between generating high

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

dispersion in sales and generating high dispersion in markups.

Since the choke quantity decreases with η/σ, by matching the same distribution of sales as in the baseline model (Figure SM.5.1 in Supplemental Materials), the homogeneous customers model assigns a much lower calibrated value to this ratio (similar to the one reported by Edmond, Midrigan, and Xu (2022)), which generates much lower dispersion in markups across firms. This is, however, not the case with our baseline model, in which firms' size distribution is also affected by the distribution of customers across firms—which is not directly related to their markups, and thus is not constrained by the choke quantity demanded by each customer.

Figure 3: Distribution of Markups: Baseline vs Homogeneous Customers Model



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

Two observations follow. First, the homogeneous customers model overestimates markups for small firms and underestimates markups for large firms relative to our baseline model. Second, and more importantly, by generating a higher markup dispersion than the homogeneous customers model, the baseline model implies much higher welfare losses due to misallocation, as we show in the next section.[28]

[28] In the Compustat data, the overall dispersion of markups is larger than the dispersion obtained from

5 Efficient Allocation

What are the implications of endogenous customer acquisition for welfare and, in particular, misallocation? In this section we start by developing a welfare decomposition result that highlights the potential sources of welfare losses or gains around the equilibrium allocation. We then characterize the efficient allocation in our economy by solving the problem of a social planner who faces the same constraints as agents do in the equilibrium. Using our decomposition result, we then decompose the equilibrium welfare losses to aggregate TFP losses (misallocation), losses from total underutilization of labor due to *aggregate* markups, and losses from different entry/exit policies for firms. Our core finding is that welfare losses from allocative efficiency in TFP are around 7.8% in consumption-equivalent terms and constitute around 57% of total welfare losses under the equilibrium allocation.

5.1. A Welfare Decomposition Result

Before stating our welfare decomposition result, we need to derive the model-consistent notions of *aggregate TFP* and *aggregate markup*. In particular, for any allocation of production inputs and demand $(l_{i,p,t}, m_{i,t}, q_{i,t})_{i\in N_t}$ across a set of operating firms N_t, we derive the aggregate TFP as the productivity implied by an aggregated production function, and the aggregate markup as the wedge between the marginal product of labor and the marginal rate of substitution between leisure and consumption.

Aggregate Production Function Let $L_{p,t}$ denote the total amount of labor allocated to production across all firms. Then, we have:

$$L_{p,t} \equiv \int_{i\in N_t} l_{i,p,t}\, di = \int_{i\in N_t} \left(\frac{C_t m_{i,t} q_{i,t}}{z_{i,t}}\right)^{\alpha^{-1}} di, \tag{5.1}$$

where the second equality follows from equating firms' demand to their supply.

Defining aggregate output as the aggregate consumption, $Y_t \equiv C_t$, and rearranging Equation (5.1), we arrive at the aggregate production function:

$$Y_t = Z_t L_{p,t}^{\alpha}, \tag{5.2}$$

either our baseline model or the homogeneous customers model. This is expected, since our model focuses only on the dispersion of markups that arises from differences in relative sales (per customer). In reality, markups or wedges in the marginal revenue of products can arise due to other distortions (see, e.g., Hsieh and Klenow, 2009) from which our model abstracts away.

where Z_t, the aggregate TFP, is given by

$$Z_t \equiv \left[\int_{i \in N_t} \left(\frac{z_{i,t}}{q_{i,t} m_{i,t}} \right)^{-\alpha^{-1}} di \right]^{-\alpha}. \tag{5.3}$$

Notice how the allocation of relative demand $(q_{i,t}, m_{i,t})_{i \in N_t}$ affects aggregate productivity. Higher relative demand means higher production relative to other firms, which decreases aggregate productivity through a lower marginal product of labor (due to decreasing returns to scale) and dispersion in demand per customer (due to imperfect substitutability of goods).

Aggregate Markup We define the aggregate markup, $\mathcal{M}_t$, as the wedge between the marginal product of aggregated production labor and the real wage W_t/P_t (which, from the household's labor supply condition, is equal to the marginal rate of substitution between leisure and consumption). Recall also that we have normalized the aggregate price to one ($P_t = 1$) so that the real wage is W_t. Formally, having derived the aggregate production function in Equation (5.2), aggregate markup is defined as

$$\mathcal{M}_t \equiv \frac{\partial Y_t / \partial L_{p,t}}{W_t} = \alpha \frac{Y_t}{W_t L_{p,t}}. \tag{5.4}$$

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

$$\mu_{i,t} \equiv \alpha \frac{p_{i,t} y_{i,t}}{W_t l_{i,p,t}}, \tag{5.5}$$

which corresponds to the equilibrium relationship between the markup and the labor share in Equation (3.16)—with the exception that here we are defining this wedge for an arbitrary allocation of inputs and prices. By combining the last two equations, we obtain the aggregate markup as the production cost-weighted average of firm-level markups (as in Edmond, Midrigan, and Xu, 2022):

$$\mathcal{M}_t = \int_{i \in N_t} \omega_{i,t} \mu_{i,t} di, \tag{5.6}$$

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

$$\omega_{i,t} \equiv \frac{W_t l_{i,p,t}}{\int_{i \in N_t} W_t l_{i,p,t}}. \tag{5.7}$$

Decomposition of Welfare Having defined aggregate TFP and markup, we obtain the following proposition that emphasizes their relevance by showing that, up to the first order, allocations affect welfare only through these and other aggregate objects.

Proposition 5. For small perturbations around the equilibrium allocation, changes in household's welfare at a given time t are given, up to a first-order approximation, by

$$\underbrace{\frac{\Delta U_t}{U_{c,t}C_t}}_{\Delta\text{Welfare (C.E.)}} \approx \underbrace{\Delta\ln(Z_t)}_{\Delta\text{TFP}} + \underbrace{\alpha(1-\mathcal{M}_t^{-1})\Delta\ln(L_{p,t})}_{\Delta\text{ from Aggregate Markup}} - \alpha\mathcal{M}_t^{-1}\left[\underbrace{\chi\frac{N_t}{L_{p,t}}\Delta\ln(N_t)}_{\Delta\text{ from Entry/Exit}} + \underbrace{\frac{L_{s,t}}{L_{p,t}}\Delta\ln(L_{s,t})}_{\Delta\text{ from Advertising}}\right], \quad (5.8)$$

where Z_t is the aggregate TFP in Equation (5.3), $\mathcal{M}_t$ is the equilibrium aggregate (cost-weighted) markup in Equation (5.6), $L_{p,t}$ and $L_{s,t}$ are the aggregate amounts of labor allocated to production and advertising, and N_t is the equilibrium measure of operating firms.

Equation (5.8) 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 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 the aggregate labor supply that is allocated to the overhead costs of operating firms, and (4) changes in the aggregate labor supply that is allocated to advertising in the equilibrium.

Proposition 5 lays out the road map for the rest of our analysis. As we move on to quantify the efficient allocation, our main objective is to measure these welfare changes under counterfactual demand allocations.

5.2. Characterization of the Social Planner's Problem

Endogenous customer acquisition creates a new channel for the relationship between relative size and markups. Not only does this new channel affect this relationship in the equilibrium, but it also defines a new Pareto frontier, since the social planner also chooses the distribution of customers across firms. This section characterizes this efficient allocation in our economy.

Given an initial distribution of firms, the social planner of this economy maximizes the household's lifetime utility by choosing (1) which incumbent firms should exit and which potential entrants should enter at each period, (2) how many customers each operating firm should get—which can be achieved either by depreciating their customer base if the firm has too many customers or by launching advertising campaigns if the firm needs to grow—and, finally, (3) how much each operating firm should produce. A formal statement of the planner's problem is included in Appendix B.8. Here, we discuss

the key properties of the solution to this problem.

Any Distribution of Customers is Attainable There are two sources of inefficiencies regarding customer acquisition and the allocation of customers in the equilibrium. First, the planner might choose to allocate customers differently across firms than the equilibrium (misallocation of customers). A second source of inefficiency is the business-stealing externality of advertising, which leads to an overuse of labor for advertising in the equilibrium.[29] To focus on the *misallocation* of customers, we shut down this second source of inefficiency by restricting the social planner to spend the same amount of aggregate labor for advertising as in the equilibrium. More precisely, the restriction shuts down Channel 4 (Δ from Advertising) in Proposition 5.

A potential concern is that shutting down the second channel in this manner might affect the distribution of customers that are attainable for the social planner and, thus, our conclusions about the misallocation of customers. This is a nontrivial concern, but it turns out that it is nonbinding: In the following lemma, we show that shutting down the second source of inefficiencies is inconsequential for the implications of customer misallocation. That is, by restricting the planner to using a certain amount of aggregate labor for advertising, we are not imposing any restriction on the planner's choices regarding the allocation of customers across firms.

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

This result follows from the advertising technology, which requires that returns to advertising are fully relative in labor allocated to posting ads. Given this result, we solve the planner's problem in two steps. First, for any set of operating firms, we characterize the optimal allocation of demand in terms of how many customers each operating firm

[29] This assumption is irrelevant for the calibration of the model because the measure of matches can be normalized in that exercise. But it is important for counterfactuals in which it is important how the total number of matches changes with firms' advertising. Assuming that the total number of matches are fixed *across* counterfactuals implies that while advertising changes the distribution of customers across firms, it does not expand the industry's total customer base as in (Drozd and Nosal, 2012, Einav, Klenow, Levin, and Murciano-Goroff, 2022). This assumption is corroborated by marketing and IO literature (Bagwell, 2007). Studies of the publishing (Garthwaite, 2014) and prescription drugs (Sinkinson and Starc, 2019) sectors indicate that advertising boosts own sales without enlarging the total size of the market. Moreover, while we have made the extreme assumption in the model that advertising has no effect on the total size of the customer base in the industry, we view this as a conservative benchmark because the planner also faces this limitation, ensuring that welfare gains come from customer reallocation only and ignoring any potential welfare gains by increasing the size of the customer base for an industry. Finally, by restricting the social planner to using the same amount of labor allocated to advertising, we ensure that our reported welfare gains are not due to the planner's using fewer resources for advertising.

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 Maximizes Aggregate TFP Here, we characterize the efficient allocation of demand for a given set of operating firms. Formally, by an allocation of demand, we mean the choice of (a) allocating customers across firms in N_t—i.e., $\mathbf{1}_{\{j\in m_{i,t}\}}, \forall j, \forall i \in N_t$—and (b) choosing the relative demand of every matched customer, $q_{i,j,t}, \forall j \in m_{i,t}, \forall i \in N_t$. Naturally, any allocation of demand must:

1. be consistent with the Kimball aggregator:

$$\int_{i\in N_t}\int_0^1 \mathbf{1}_{\{j\in m_{i,t}\}}\Upsilon(q_{i,j,t})\,dj\,di = 1, \tag{5.9}$$

2. respect the constraint for the total number of matches:

$$\int_{i\in N_t}\int_0^1 \mathbf{1}_{\{j\in m_{i,t}\}}\,dj\,di \leq 1. \tag{5.10}$$

Proposition 6. Fix a choice for the set of operating firms, N_t. Then, the efficient allocation of demand is the one that maximizes aggregate TFP subject to constraints in Equations (5.9) and (5.10). This allocation equalizes demand per *matched* customer across all firms, and matches more customers to more productive firms:

$$q_{i,j,t} = 1, \forall j \in m^*_{i,t}, \quad m^*_{i,t} = \frac{z_{i,t}^{\frac{1}{1-\alpha}}}{\int_{i\in N_t} z_{i,t}^{\frac{1}{1-\alpha}}\,di}. \tag{5.11}$$

Proposition 6 shows that the planner equalizes the level of consumption per customer across all firms. Furthermore, to capitalize on the higher efficiency of more productive firms, the planner gives them *more customers*.[30] This is in contrast to the equilibrium, in which more productive firms have both higher sales per customer and more customers than other firms (but potentially fewer customers than what is efficient).

Our result in Proposition 6 is also at odds with the trade-off the social planner faces in the homogenous customers model, in which all firms are assumed to serve the representative consumer. On the one hand, the social planner would like more productive firms to produce more in order to increase *aggregate consumption* (i.e., to equalize the

[30] Note that the allocation of customers does not depend on the initial distribution of matches. This follows from Lemma 2. Since the implementation cost of all distributions is the same for the planner, we can assume without loss of generality that the planner exercises the free disposal of matches at the beginning of every period and re-matches all customers based on firms' new productivities. This might lead some firms to lose customers beyond those exogenously depreciated. We note that the same option is also available to firms in the competitive equilibrium, but they choose not to exercise it (see the proof of Proposition 2).

marginal product of inputs across firms). On the other hand, since demand comes from the intensive margin in those models, instructing more productive firms to produce more creates dispersion in *relative consumption* across varieties, which is costly due to the weak substitutability of goods. Therefore, the social planner has to balance these two opposing forces when choosing the optimal allocation of inputs.

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

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

Social Value of a Firm In characterizing the efficient allocation, we show that to make entry or exit decisions, the social planner considers the lifetime valuation of a firm for the welfare of the representative household, which we denote by the social value of a firm. In order to compare how the social planner values individual firms relative to the equilibrium, the following proposition characterizes the social value of firms under the efficient allocation.

Proposition 7. The social value of any firm at any given time depends only on its current productivity and is given by

$$v_t^*(z_{i,t}) = \max_{\substack{\{\mathbf{1}_{i,\tau}, m_{i,\tau}\} \\ \{q_{i,\tau}, l_{i,p,\tau}\}}_{\tau \geq t}} \mathbb{E}_t \sum_{\tau=t}^{\infty} (\beta \nu)^{\tau - t} \left(\prod_{h=t}^{\tau} \mathbf{1}_{i,h} \right) \left(\frac{C_\tau^*}{C_t^*} \right)^{-\gamma} \left(D_\tau^* \frac{\Upsilon(q_{i,\tau})}{q_{i,\tau}} y_{i,\tau} - W_t^* (l_{i,p,\tau} + \chi) - C_\tau^* (D_\tau^* - 1) m_{i,\tau} \right) \tag{5.12}$$

$$s.t. \quad y_{i,\tau} = m_{i,\tau} q_{i,\tau} C_\tau^* = z_{i,\tau} l_{i,p,\tau}^{\alpha}, \tag{5.13}$$

where $(C_\tau^*, D_\tau^*, W_\tau^*)_{\tau \geq t}$ are the aggregate consumption, aggregate demand index, and decentralized wage under the planner's allocation.

The first observation is that the social value of a firm does not depend on its initial customer base at a given time t. This follows from Lemma 2: Since any distribution

of matches is attainable for the social planner using the matching technology of the economy, the efficient allocation is not restricted by the initial distribution of matches at any given time. Accordingly, the social planner can choose the optimal number of customers for any firm given their current productivity. If a firm has too many customers, the planner simply exercises the free disposal condition of matches (i.e., choose $\delta_{i,t} \geq \delta$), and if a firm has too few customers, the planner allocates more ads for that firm in relative terms to attain the optimal number of matches.

Moreover, the expression in Equation (5.12) clarifies how the planner's allocation for a given firm differs from its equilibrium counterpart in Equation (3.13). There are three notable differences. First, the social planner maximizes a firm's value based on the *average consumer surplus* ($\Upsilon(q)/q$). In contrast, in the equilibrium, the revenue firms receive is based on the *marginal valuation* ($\Upsilon'(q)$). Second, the advertising labor cost of allocating customers does not appear in this value because aggregate advertising labor is a sunk cost for the planner and, at the margin of keeping a firm, it does not affect her decision. Third, there is a new term ($C_\tau^*(D_\tau^* - 1)m_{i,\tau}$) that captures the externalities of allocating customers across firms, which is endogenous to the planner but neglected by individual firms. This externality arises from the fact that, fixing the total number of customers in the economy, if one firm is being matched to a particular customer, then another firm must lose one.[31]

5.3. Quantifying the Efficient Allocation

This section presents the differences between the equilibrium and efficient allocations and, in particular, quantifies the welfare gains under the efficient allocation of demand. In doing so, we also revisit our exante decomposition of welfare in Proposition 5 and quantify the contribution of each of the three channels (TFP, aggregate markups, and overhead costs). To isolate the role of endogenous customer acquisition, we conclude this section by repeating the same exercise for counterfactual values of ϕ, the parameter that governs returns to scale in customer acquisition for firms.

Welfare Gains We start by quantifying the three channels of welfare gains from Proposition 5 in our calibrated model per Proposition 5 (recall that we shut down welfare gains from advertising labor—Channel 4—by restricting the social planner to using the same amount of labor).

[31] Note that this effect would still exist even if advertising increases the total number of customers because, conditional on having spent on advertising labor and creating a match, the planner still has to decide which firm to allocate the match to.

$$\underbrace{\frac{\Delta U_t}{U_{c,t}C_t}}_{\Delta\text{Welfare (C.E.) = 13.6\%}} \approx \underbrace{\Delta \ln(Z_t)}_{\text{TFP gains = 10.8\%}} \underbrace{-\alpha \mathscr{M}_t^{-1} \chi \frac{N_t}{L_{p,t}} \Delta \ln(N_t)}_{\text{Gains from Entry/Exit = 1.6\%}} \underbrace{+\alpha(1-\mathscr{M}_t^{-1})\Delta \ln(L_{p,t})}_{\text{Losses from Underutilization of Labor = 0.78\%}} . \tag{5.14}$$

There are two main takeaways from this decomposition: (1) the consumption-equivalent welfare gains of the household under the efficient allocation are substantial and quantified at 13.6%, and (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 also achieves 1.6% higher welfare by reducing the amount of labor allocated to 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 6 presents the implied changes in other quantities that arise from these gains. As a result of higher TFP and higher production labor, the output is 14.6% higher under the efficient allocation, even though the number of firms is 11.3% lower. This higher production with fewer firms is made possible by the fact that the concentration of sales among the top 5% largest firms is 39.2% larger than in the equilibrium. In addition to the baseline model, Table 6 presents similar results for two counterfactual values of ϕ. In the remainder of this section, we dissect these changes and study the underlying forces that shape these gains.

Implications for Misallocation By explicitly modeling the extensive margin of demand in a way consistent with the data, we find large TFP gains of 10.8%. This shows that efficiency losses from the misallocation of customers are large and go well beyond the social costs of markups (for instance, Edmond, Midrigan, and Xu (2022) estimate the efficiency losses from markups to be around 0.8% to 1.8%).

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

Table 6: 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.

allocations as

$$\underbrace{\ln\left(\frac{Z(N_t^*,\mathscr{A}_t^*)}{Z(N_t,\mathscr{A}_t)}\right)}_{\Delta\text{ TFP} = 10.8\%} = \underbrace{\ln\left(\frac{Z(N_t,\mathscr{A}_t^*)}{Z(N_t,\mathscr{A}_t)}\right)}_{\Delta\text{ Allocative Efficiency} = 7.8\%} + \underbrace{\ln\left(\frac{Z(N_t^*,\mathscr{A}_t^*)}{Z(N_t,\mathscr{A}_t^*)}\right)}_{\Delta\text{ Entry/Exit Efficiency} = 3.0\%} . \tag{5.15}$$

The first term on the right-hand side of Equation (5.15) shows that, keeping the set of operating firms fixed, almost 75% of the efficiency gains under the planner's allocations are due to allocative efficiency gains. This is the most important consequence of endogenous customer acquisition: Having the ability to reallocate customers across firms, the planner shifts the distribution of customers toward the top of the productivity distribution, and hence is able to allocate higher amounts of production labor to them.

The extensive margin of demand is the key to the high TFP gains in our model: Without the extensive margin, the planner can only achieve a higher aggregate productivity by shifting demand toward more productive firms on the intensive margin. However, since varieties are weak substitutes, distorting the distribution of relative demand is costly. These costs are even higher when demand is more elastic at higher quantities (as with Kimball preferences or any semi-kinked demand system).

However, in this model, the efficient allocation is not restricted by the weak substitutability margin: Completely equalizing relative consumption across individuals

($q^*_{i,t} = 1$), the efficient allocation achieves much higher aggregate productivity by simply allocating *more customers* to more productive firms ($m^*_{i,t} \propto z_{i,t}^{\frac{1}{1-\alpha}}$)—as seen in Figure SM.4.12 in Supplemental Materials, which shows a comparison of the allocation of customers between the equilibrium and the efficient allocation. As a consequence of this reallocation of customers, the concentration of sales among the top 5% of firms increases by 39.2%. This higher concentration of customers among more productive firms is efficient to the point that marginal costs of production are equalized across all firms.

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

Finally, while the optimal allocation of resources accounts for around 75% of the change in aggregate TFP, the remaining 25% is explained by sheer compositional changes in the distribution of productivity, i.e., technological change. Under the efficient allocation, the planner is more selective in allowing firms to enter and chooses a higher productivity cutoff for the entry and exit of firms. A more selective policy increases TFP because it increases the average productivity of firms that operate in the economy and implies fewer operating firms.

The Optimal Number of Firms We start by reviewing the usual costs and benefits of having more firms in the economy and then discuss the new mechanism that comes into play in our model. In conventional models, the optimal number of firms is affected by the interaction of three forces: decreasing returns to scale, love of variety, and aggregate overhead costs. On one side, with decreasing returns to scale at the firm level, having more firms increases the aggregate efficiency by dividing resources across a larger number of firms. Moreover, with love-of-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 experiences higher productivity.[32] While these forces form the benefits of a

[32] Both of these forces can be summarized by the following simple example inspired by Edmond, Midrigan, and Xu (2022). Consider an economy with N firms indexed by i, where every firm produces with $y_i = l_i^{\alpha}$ and aggregate output is given by a CES aggregator, $Y = [\int_0^N y_i^{\theta-1} di]^{\theta}$. For a given amount of aggregate labor, L, every firm gets to produce $y_i = (L/N)^{\alpha}$ and the aggregate output is given by $Y = N^{\theta-\alpha} L^{\alpha}$. Now if we shut down love of variety ($\theta = 1$), productivity is $N^{1-\alpha}$, which increases with N. If we shut down decreasing returns to scale ($\alpha = 1$), productivity is $N^{\theta-1}$, which indicates higher productivity due to love of variety with

higher number of firms, the cost is usually modeled either as a fixed entry cost for every firm or, as we model here, a stream of overhead costs over time, both of which lead to an optimal finite measure of firms in the equilibrium.

Moreover, our model has an additional force that arises from the allocation of customers across firms. Since the number of customers is fixed, a higher concentration of customers at the top comes at the cost of fewer customers at the bottom of the productivity distribution, which in turn reduces the social value of such firms (as shown in Proposition 7). Hence, with this additional instrument, our planner achieves a higher TFP without having to pay for the overhead costs of more firms, which leads to a lower number of firms in the efficient allocation and increases the welfare of the household by 1.6%, as shown in Equation (5.3).

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

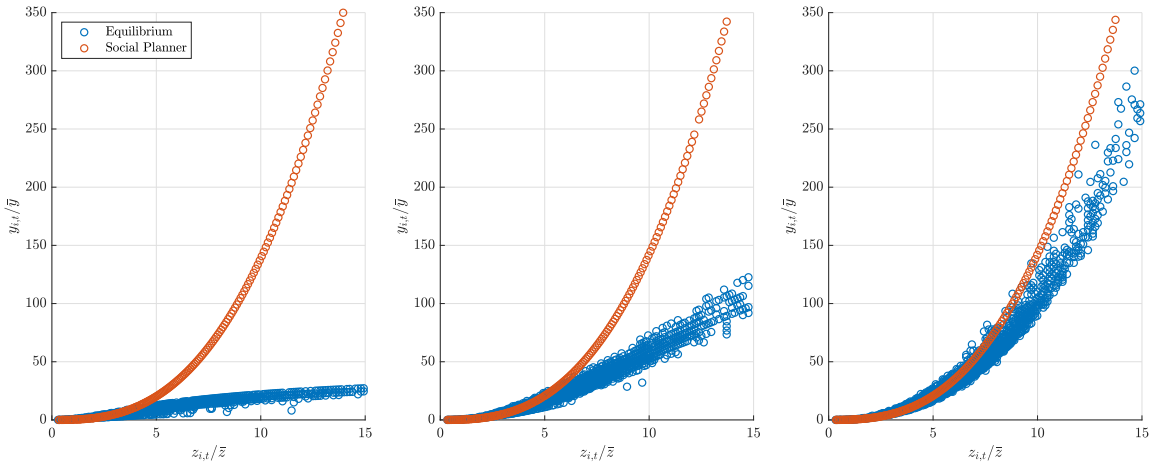
The Role of Returns to Scale in Marketing While for the planner the only relevant margin in allocating customers is the returns to scale in production, the efficiency gains from the reallocation of customers depend on the returns to scale for customer acquisition, ϕ. Larger returns to scale in customer acquisition would imply that more productive firms would invest more in customer acquisition, which is desirable from the perspective of the efficient allocation. Figure 4 shows the scatter plot of firms' productivity and output for both the equilibrium and social planner's allocation and for three different values of ϕ (a low value, the calibrated value, and a high value). The figure shows that with a larger ϕ the equilibrium allocation of customers is closer to that of the planner.[33]

Finally, the $\phi = 0.25$ and $\phi = 0.75$ columns of Table 6 show how the allocation of customers is solely responsible for the large efficiency gains under the planner's allocation.

larger N.

[33] In fact, for the special case of $\delta = 1$, the allocation of customers in the equilibrium coincides with the efficient allocation when $\phi \to 1$.

Figure 4: Allocation of Output: Equilibrium vs. Efficient Allocation



(a) Low ϕ (b) Baseline (c) High ϕ

Notes: This figure shows a scatter plot between relative productivity $z_{i,t}/\bar{z}$ and relative output $y_{i,t}/\bar{y}$ for both the equilibrium and the social planner's allocation. We present three plots by varying the value of ϕ, and 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.

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, and as a result charge lower markups. Hence, aggregate TFP, output, and concentration increase while the aggregate markup decreases and the economy gets closer to the efficient allocation.

6 Conclusion

In this paper, we revisit the role of the extensive and intensive margins of demand in firms' market share and market power. Using a dataset that merges information from the consumer and the producer sides, we document that while firms' sales grow mainly through acquiring more customers, their market power is only correlated with their average sales per customer.

Guided by these empirical findings, we develop and quantify a model that micro-founds the relationship between market power and concentration in the extensive and intensive margins. In our model, while firms hold market power over each customer,

the total number of customers acts as a demand shifter. The model provides a new perspective on the relationship between relative 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. We also 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 market power through concentration and profits. In particular, our model highlights a new unintended consequence of policies that target only firms' market power. In our model, although market power is distortionary, it compensates more productive firms for their investment in customer acquisition and improves the allocation of customers. Thus, policies that target larger firms disproportionately may have adverse effects through the misallocation of customers. In particular, 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, M., O. ITSKHOKI, AND J. KONINGS (2019): "International Shocks, Variable Markups, and Domestic Prices," *Review of Economic Studies*, 86(6), 2356–2402.

ANDREWS, I., M. GENTZKOW, AND J. M. SHAPIRO (2017): "Measuring the Sensitivity of Parameter Estimates to Estimation Moments," *Quarterly Journal of Economics*, 132(4), 1553–1592.

ARGENTE, D., D. FITZGERALD, S. MOREIRA, AND A. PRIOLO (2021): "How Do Firms Build Market Share?," Manuscript.

ARGENTE, D., M. LEE, AND S. MOREIRA (2018): "Innovation and Product Reallocation in the Great Recession," *Journal of Monetary Economics*, 93, 1–20.

——— (2019): "The Life Cycle of Products: Evidence and Implications," Manuscript.

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

ARNOUD, A., F. GUVENEN, AND T. KLEINEBERG (2019): "Benchmarking Global Optimizers," Manuscript.

ATKESON, A., AND A. BURSTEIN (2008): "Pricing-to-Market, Trade Costs, and International Relative Prices," *American Economic Review*, 98(5), 1998–2031.

BAGWELL, K. (2007): "The economic analysis of advertising," *Handbook of Industrial Organization*, 3, 1701–1844.

BAQAEE, D. R., AND E. FARHI (2019): "Productivity and Misallocation in General Equilibrium*," *Quarterly Journal of Economics*, 135(1), 105–163.

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

BOLTON, R. N., P. K. KANNAN, AND M. D. BRAMLETT (2000): "Implications of Loyalty Program Membership and Service Experiences for Customer Retention and value," *Journal of the Academy of Marketing Science*, 28(1), 95–108.

BOND, S., A. HASHEMI, G. KAPLAN, AND P. ZOCH (2021): "Some unpleasant markup arithmetic: Production function elasticities and their estimation from production data," *Journal of Monetary Economics*.

BORNSTEIN, G. (2021): "Entry and Profits in an Aging Economy: The Role of Consumer Inertia," Mimeo.

BORNSTEIN, G., AND A. PETER (2022): "Nonlinear Pricing and Misallocation," Mimeo.

BURSTEIN, A., V. M. CARVALHO, AND B. GRASSI (2020): "Bottom-up Markup Fluctuations," Manuscript.

CAVENAILE, L., M. A. CELIK, J. PERLA, AND P. ROLDAN-BLANCO (2023): "A model of product awareness and industry life cycles," Manuscript.

CLEMENTI, G. L., AND B. PALAZZO (2016): "Entry, Exit, Firm Dynamics, and Aggregate Fluctuations," *American Economic Journal: Macroeconomics*, 8(3), 1–41.

DAVIS, S. J., AND J. HALTIWANGER (1992): "Gross Job Creation, Gross Job Destruction, and Employment Reallocation," *Quarterly Journal of Economics*, 107(3), 819–863.

DAVIS, S. J., J. C. HALTIWANGER, S. SCHUH, ET AL. (1998): "Job Creation and Destruction," *MIT Press Books*, 1.

DE LOECKER, J., J. EECKHOUT, AND G. UNGER (2020): "The Rise of Market Power and the MacRoeconomic Implications," *Quarterly Journal of Economics*, 135(2), 561–644.

DOTSEY, M., AND R. G. KING (2005): "Implications of State-Dependent Pricing for Dynamic MacRoeconomic Models," *Journal of Monetary Economics*, 52(1), 213–242.

DROZD, L. A., AND J. B. NOSAL (2012): "Understanding International Prices: Customers as Capital," *American Economic Review*, 102(1), 364–395.

EDMOND, C., V. MIDRIGAN, AND D. Y. XU (2022): "How Costly Are Markups?," Manuscript.

EINAV, L., P. J. KLENOW, J. D. LEVIN, AND R. MURCIANO-GOROFF (2022): "Customers and Retail Growth," Manuscript.

ELSBY, M. W., AND R. MICHAELS (2013): "Marginal Jobs, Heterogeneous Firms, and Unemployment Flows," *American Economic Journal: Macroeconomics*, 5(1), 1–48.

FITZGERALD, D., S. HALLER, AND Y. YEDID-LEVI (2016): "How Exporters Grow," *National Bureau of Economic Research Working Paper Series*.

FOSTER, L., J. HALTIWANGER, AND C. SYVERSON (2015): "The Slow Growth of New Plants: Learning About Demand?," *Economica*, 83(329), 91–129.

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

GOPINATH, G., P.-O. GOURINCHAS, C.-T. HSIEH, AND N. LI (2011): "International Prices, Costs and Mark-up differences," *American Economic Review*, 101(6), 2450–86.

GOPINATH, G., AND O. ITSKHOKI (2010): "Frequency of Price Adjustment and Pass-Through," *Quarterly Journal of Economics*, 125(2), 675–727.

HALTIWANGER, J., R. S. JARMIN, AND J. MIRANDA (2013): "Who Creates Jobs? Small Versus Large Versus Young," *Review of Economics and Statistics*, 95(2), 347–361.

HOPENHAYN, H. A. (1992): "Entry, Exit, and Firm Dynamics in Long Run Equilibrium," *Econometrica*, pp. 1127–1150.

HOTTMAN, C. J., S. J. REDDING, AND D. E. WEINSTEIN (2016): "Quantifying the Sources of Firm Heterogeneity," *Quarterly Journal of Economics*, 131(3), 1291–1364.

HSIEH, C.-T., AND P. J. KLENOW (2009): "Misallocation and Manufacturing TFP in China and India," *Quarterly Journal of Economics*, 124(4), 1403–1448.

KAPLAN, G., AND P. ZOCH (2020): "Markups, Labor Market Inequality and the Nature of Work," *National Bureau of Economic Research Working Paper Series*.

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

KLENOW, P. J., AND J. L. WILLIS (2016): "Real Rigidities and Nominal Price Changes," *Economica*, 83(331), 443–472.

LEE, Y., AND T. MUKOYAMA (2015): "Productivity and Employment Dynamics of US Manufacturing Plants," *Economics Letters*, 136, 190–193.

MITTAL, V., AND W. A. KAMAKURA (2001): "Satisfaction, Repurchase Intent, and Repurchase Behavior: Investigating the Moderating Effect of customer characteristics," *Journal of Marketing Research*, 38(1), 131–142.

NAKAMURA, E., AND D. ZEROM (2010): "Accounting for Incomplete Pass-Through," *Review of Economic Studies*, 77(3), 1192–1230.

NEIMAN, B., AND J. S. VAVRA (2019): "The Rise of Niche Consumption," Manuscript.

OTTONELLO, P., AND T. WINBERRY (2020): "Financial heterogeneity and the investment channel of monetary policy," *Econometrica*, 88(6), 2473–2502.

PETERS, M. (2020): "Heterogeneous markups, growth, and endogenous misallocation," *Econometrica*, 88(5), 2037–2073.

PHELPS, E. S., AND S. G. WINTER (1970): "Optimal Price Policy Under Atomistic Competition," *Microeconomic foundations of employment and inflation theory*, pp. 309–337.

RESTUCCIA, D., AND R. ROGERSON (2008): "Policy Distortions and Aggregate Productivity With Heterogeneous Establishments," *Review of Economic Dynamics*, 11(4), 707–720.

ROTEMBERG, J. J., AND M. WOODFORD (1999): "The Cyclical Behavior of Prices and Costs," *Handbook of Macroeconomics*, 1, 1051–1135.

SINKINSON, M., AND A. STARC (2019): "Ask your doctor? Direct-to-consumer advertising of pharmaceuticals," *Review of Economic Studies*, 86(2), 836–881.

STROEBEL, J., AND J. VAVRA (2019): "House Prices, Local Demand, and Retail Prices," *Journal of Political Economy*, 127(3), 1391–1436.

TRAINA, J. (2019): "Is Aggregate Market Power Increasing? Production Trends Using Financial Statements," Manuscript.

WASI, N., AND A. FLAAEN (2015): "Record Linkage Using Stata: Preprocessing, Linking, and Reviewing Utilities," *The Stata Journal*, 15(3), 672–697.

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