Measuring Customer Churn and Interconnectedness

Scott R. Baker* Brian Baugh† Marco Sammon‡

May 2020

Abstract

This paper demonstrates that it is possible to construct accurate pictures of firm revenue, growth, geographic dispersion, and customer base characteristics using an increasingly accessible class of consumer financial transaction data. We develop two new measures which characterize firms' customer bases: the rate of churn in a firm's customer base and a metric of the pairwise similarity between firms' customer bases. We show that these measures provide important insights into the behavior of both real firm decisions and firm asset prices. Rates of customer churn affect the level and volatility of firm-level investment, markups, and profits. Churn also affects how quickly firms respond to shocks in the value of their growth options (i.e. Tobin's Q). Moreover, high churn firms tended to face steeper declines in consumer spending during the recent COVID-19 outbreak. Similarity between firms' customer bases highlights one under-explored type of predictability among stock returns – we demonstrate that significant alpha can be generated using a trading strategy that exploits our index of customer base similarity across firms.

JEL Classification: D22, E22, G32, L11

Keywords: customer base, customers, transaction data, customer churn, customer similarity

*Northwestern University, Kellogg School of Management s-baker@kellogg.northwestern.edu.

†University of Nebraska bbaugh2@unl.edu.

‡Northwestern University, Kellogg School of Management mcsammon@gmail.com.

1 Introduction

For firms throughout the economy, customer bases represent one of the most important assets as well as one of the biggest sources of risk. Understanding the composition and behavior of customer bases can be crucial for understanding both firm asset prices and real economic decisions. While some firms interact with an ever-changing set of customers, other firms build a durable customer base that adjusts only slowly to firm-level shocks. Differences in the durability of customer bases across firms is likely to influence firm behavior. For example, firms with stable customer bases may be able to charge higher markups than firms with more transient customer bases.

Moreover, the specific customers making up a customer base can represent a source of risk for firms. Shocks to the income of customer bases can drive revenue growth, volatility, and correlations across firms. Firms both across and within industry may have substantial overlaps in customer bases, yielding a new vector for demand shocks (e.g. income shocks among customers) to affect multiple firms at once.

Despite the implications that customer bases and customer networks may have for firm-level outcomes, it has been difficult to obtain systematic data about firm customer bases for any particular firm, much less a full network of firms and linkages through their customer bases. In this paper, we address this shortfall in data by utilizing financial transaction data to gain insights into the composition, characteristics, and time-series variation in firm-level customer bases.

Numerous researchers in the United States and around the world have begun to gain access to detailed financial transaction data and conduct research focusing mainly on questions relating to household decision-making: consumption and behavioral finance, microeconomic foundations of aggregate shocks, policy analysis, and individual equity market behavior.[1] While financial transaction data has already transformed these fields related to households and consumers, this paper shows that this class of data has substantial utility when applied to research regarding consumer-facing *firms*, as well.

In part, this paper works to demonstrate that financial transaction data can be applied to many more questions and fields than has currently been the case. To our knowledge, Agarwal et al.

[1] Some research utilizing financial transaction data has used sources from Mexico (Bachas et al., 2019), Singapore (Agarwal and Qian, 2014), Brazil (Medina, 2020), Turkey (Aydin, 2019), Germany (Baker et al., 2020), Iceland (Olafsson and Pagel, 2018), and the United States (Ganong and Noel, 2019).

(2020) is perhaps the only other paper that takes this view using transaction-level data. They demonstrate that disaggregated spending data can provide high quality signals about consumer demand and equity prices for consumer-facing firms. They also find some indications that the types of customers that patronize a given retailer predicts the equity market response to spending shocks at that retailer.

In this vein, we show that using individual financial transaction data to understand firm customer bases can allow for unique insights into industry-level concentration and competition, the real effects of marketing and advertising on customer behavior, how regional shocks propagate through the economy, and how customer bases respond to mergers or financial distress within firms. We develop and validate several metrics relating to firms' customer bases and demonstrate two applications in which such metrics yield insights into both the real economic outcomes of firms as well as the extent to which their stock prices may co-move.

In this paper, we utilize a transaction-level database that covers debit and credit card spending across approximately two million American users to gain insights about the firms that they patronize. One limitation of utilizing this type of data is that we are unable to analyze firms across all industries. Firms in industries like business services, manufacturing, wholesale retail, and government do not interface directly with the household sector and thus will not be the counter-party to any consumer credit or debit transactions.

As a consequence, our high quality window into the customer bases – and source of revenue – of firms is limited to the set of consumer facing firms like grocery stores, restaurants, retailers, utilities, airlines, hotels, and many online services. In this paper, we match to 558 firms, 428 of whom are publicly traded and 130 of which are private. While we are limited to a subset of firms in the economy, for this set of firms we can observe the distribution of firm's revenue as well as data on each individual customer and the other firms at which that customer shops. Finally, we can observe how these customer base characteristics change over time for each matched firm.

Our first task is to demonstrate that the picture we obtain of a given firm from its customer base yields accurate information. To do so, we compare financial and geographic characteristics for the matched firms within our data to similar data obtained from external sources. In particular, we show that, using our financial transaction data, we can accurately predict firm revenue levels and growth rates when compared to data from Compustat. Second, we compare the geographic

distribution of revenue to firms that we observe in our data and compare to an external measure of the geographic spread of firms' establishments. We find a close correlation between the two, demonstrating that we can obtain clear signals of not only aggregate revenue and growth rates, but also the geographic dispersion of firm revenue. Finally, we compare elements of the observed customer bases, like average levels of income. We find that the observed average income of customer bases predicts firm quality and prices obtained from Yelp.com.

With our validated data, we turn to generating some stylized facts about firm customer bases and how they shift over time. We first construct an index of the rate of churn in a firm's customer base over time. We also design a metric of the pairwise similarity between two firms' customer bases.

These two new measures are quite distinct from existing firm-level data. For instance, while there are other measures of similarity across firms, these typically measure distance in terms of product attributes or competitive distance (eg. Hoberg and Phillips (2010)). Our measure of customer-base similarity captures a crucial element of correlated risks and similarity from the demand side. Moreover, measuring customer churn directly provides a cleaner and more direct indicator of customer loyalty and stability than existing proxies like SG&A spending. Researchers working with a given firm's customer data may be able to calculate customer churn for a single firm, but lack the cross-sectional reach that transaction data provides.

We show that both of these new measures provide important insights into the behavior of both real firm decisions and firm asset prices. First, we find that rates of customer churn affect the level and volatility of firm-level investment, markups, and profits, as well as how quickly firms respond to shocks. Second, we find that customer base similarity measures are a source of correlated demand shocks for retailers' goods – if customers face an income shock, all firms they frequent may be affected – and can lead to predictable asset price co-movements among customer-linked firms.

In our first application, we use our measure of customer base churn to test predictions of a model in which a firm's customer base acts as a state variable – firms invest in customer acquisition and retention and thus the customer base is sticky and adjusts only slowly over time. This model acts as one possible foundation of an adjustment cost model of firm investment and yields a number

of predictions about real firm outcomes.[2] In particular, such a model would predict that firms with lower levels of customer base churn would have higher rates of profitability, investment, and markups and would also respond more slowly to shocks to the firm over time.

In earlier work, Gourio and Rudanko (2014) test for the presence of a relationship between firm's customer bases and these financial outcomes among publicly traded U.S. firms. Without a direct measure of customer base churn, they rely on Selling, General and Administrative Expenses (SG&A) or advertising expenses across industries to proxy for the level of 'customer capital' that an industry has: the higher the ratio of SG&A expenses to sales, the higher the predicted level of customer capital would be. Using our direct measure of customer base churn, we are able to confirm the presence of these relationships at an individual firm-level. We also demonstrate that, while SG&A spending tends to correlate with customer capital in most industries, it is a poor measure of customer capital, measured more directly by customer base churn, within the retail and restaurant sector.

We then test whether our measure of customer base churn predicts consumer responses to the COVID-19 outbreak. We find that firms in the top quartile of customer churn experience declines in consumer spending that are at least 10 percentage points larger in magnitude than firms in the bottom quartile of customer churn.

Our second application explores the potential to better understand correlations in the demand-side risk faced by firms – firms with overlapping customer bases are subject to correlated demand fluctuations following shocks to their customer bases. If these customer base similarities are unobserved or ignored by market participants, asset prices may only respond with a lag to information about demand shocks.

In a related test of firm linkages generating predictable asset price changes, Cohen and Frazzini (2008) examine the predictability of stock returns among firms who act as supplier-customer pair. They note that markets often do not fully incorporate information regarding the likely future profitability of the supplier following news releases (measured as earnings report surprises) by the customer.

In this paper, we perform a similar analysis, testing whether the stock returns of firms tend to predict future stock returns of firms with similar customer bases. We find strong evidence that this

[2] See e.g. Christiano et al. (2005), Eberly et al. (2012).

relationship holds for both stock returns and earnings announcement surprises. Moreover, after a surprising earnings release by firm i, future analyst accuracy for similar firms $j \in J$ are unaffected. That is, analysts seem to be unaware of (or unable to measure) these customer base similarities, responding only with a substantial lag and allowing for the possibility of predictable returns among these similar firms. We find that a trading strategy exploiting our measure of customer base similarity can generate annual alpha of approximately 5%.

Overall, this paper demonstrates the utility of detailed firm-level data regarding customer bases and customer networks. While individual-level financial transaction data has been utilized to great effect in the study of household behavior, its use in fields like industrial organization, marketing, asset pricing, and corporate finance has been relatively minimal. We demonstrate that this type of data can generate new insights into the heterogeneous real economic outcomes of firms and also into previously unobservable firm-firm networks that predict co-movement in asset prices across firms.

The rest of the paper is organized as follows. Section 2 describes our data and the procedures taken to match credit and debit card transactions to firms. Section 3 presents some stylized facts about customer bases and details our measures of customer base churn and customer base similarity across firms. Section 4 presents evidence that we are observing accurate pictures of firm customer bases and their characteristics. Sections 5 and 6 lay out two applications in which we demonstrate that understanding firm's customer bases can yield important insights about asset prices and real economic decisions taken by firms. Finally Section 7 concludes.

2 Data

2.1 Transaction-Level Linked-Account Data

Online aggregation of financial accounts is a popular service that allows users to easily monitor financial activities from across multiple financial institutions using a single web-page or smartphone app. Account aggregation services often allow features such as budgeting, expense tracking, etc. Dozens of companies currently provide such services and our data comes from one of the largest of these firms.

Once a user initially signs up for the free service, they are given the opportunity to provide

the service with user-names and passwords to a variety of financial accounts (checking, savings, credit card, brokerage, retirement, mortgage, student loan, etc.) from any financial institution, though our particular data is limited to bank and credit card accounts. After signing up, the service automatically and regularly pulls data from the user's financial institutions. The data contains transaction-level data similar to those typically found on monthly bank or credit card statements, containing the amount, date, and description of each transaction. The full dataset contains 2.7 million users from 2010 to 2015 and, though the sample grows over time, there is very little attrition in our sample.

Our data is not a random sample of the population, but it appears to be widely representative, with some exceptions. In Baugh et al. (2018) and Baugh et al. (2020), the authors illustrate the income distribution of users in this database relative to the U.S. Census. While the raw sample differs from the true income distribution in the United States, the sample covers users with a wide range of incomes rather than solely identifying users of a particular income group. We also find that users in our sample are well dispersed geographically, though we have higher concentrations of users in the states of California, New York, and Texas relative to true population distributions. However, dropping members from any given state (e.g. overrepresented states) or applying other weighting strategies, does not substantially impact our results. Similarly, excluding users in the top or bottom deciles of income has little impact on our empirical results.

One challenge with working with aggregator data is the measurement of key variables, such as income and consumption. Our ability to accurately measure income, for example, depends on whether a user has linked the checking or savings account that receives their direct deposit income. If we observe no income in linked checking or savings accounts, it is impossible for us to determine whether the user truly has zero income or is simply receiving income in an unlinked account. To mitigate this concern, restrict our analysis to the subset of users for whom we observe income flowing to their checking or savings accounts. Specifically, we exclude from our analysis any user with less than $500 per month in income. In Figure 1, we compare the income distribution of remaining users to that of the U.S. Census in 2014, our last full calendar-year of aggregator data. As shown, our dataset has less population than the U.S. Census when looking at household incomes of less than $10,000 per year. However, there is considerable overlap in the remaining income distribution.

To address the issue of unobserved consumer spending due to unlinked credit cards, we remove any user who makes excessive credit card payments from the bank account relative to observed spending in the credit card account. Specifically, we remove from the sample any user that, over our entire sample period, spends twice as much on credit card payments than observed credit card spending. This has the effect of removing users which we believe have substantial amounts of spending that we do not observe transactions for. A similar restriction could be made for regular transfers from unlinked checking accounts, though these are comparatively rare as Americans tend to have a range of credit cards but generally only one or two checking accounts.

Recent work has also utilized similar transaction-based sources to make inferences about the financial habits of the broader population. For instance, Baker (2018), and Kueng (2018) also utilize data from an online personal finance platform. They perform a multitude of validity tests comparing to data sources such as Census Retail Sales, home price data from Zillow, the Survey of Consumer Finance, and the Consumer Expenditure Survey. They find a close parallel between household-level financial behaviors and distributions in these sources relative to those found among users of the online platform. That is, conditional on basic demographic types, selection into the online platform did not predict differential financial behavior or characteristics.

Ganong and Noel (2019) and Olafsson and Pagel (2018) perform similar exercises using data taken from JPMorgan Chase and a financial services app covering the population of Iceland, respectively. Across a range of financial indicators, they find strong evidence of external validity of their results using their sample population. Such results point to the fact that, while these types of bank-derived sources will mechanically exclude financial activity by the unbanked, transaction-level financial data can produce accurate portrayals of aggregate economic activity and household behavior.

2.2 Matching Procedure

2.2.1 Transaction Description Cleaning

We begin our analysis by working to match credit and debit card transactions that we observe to firms that we can then link to time-varying firm characteristics and financial performance. The initial universe of transaction description strings is made up of about 25 million unique strings. This

reflects not only a large number of unique firms, but also differences in description strings within firm driven by things like numeric transaction descriptions (e.g. 'txn: 491349'), establishment locations (e.g. 'walmart super center lancaster'), and how different credit and debit cards include or exclude punctuation.

Our first step is to reduce this count of unique strings by removing capitalization, numeric characters, punctuation, and common components (e.g. 'inc'). We are then left with approximately 1.5 million unique cleaned strings. Table 1 displays some samples of the transaction descriptions in our dataset. For each of these unique cleaned descriptions, we display the number of times that transaction is observed in our data from 2010-2015, the average transaction amount, the fraction of transactions that are debited from an account (instead of credited), and the fraction of transactions that are similar to a previous transaction to that description within a user.

As is easily seen, some transactions are much more commonly observed than others. This reflects both the relative size of retailers but also the degree to which a given retailer has different descriptions for different locations or types of transactions. For instance, we estimate that Walmart Inc. (and its subsidiary Sam's Club) is the merchant referred to in approximately 15,000 unique description strings that span different types of Walmart stores (e.g. 'Neighborhood Market', 'Super Center'), different locations, and differences in whether debit or credit cards were used.

2.2.2 Firm Selection and Matching

Given our sample of 1.5 million unique strings, we then set out to develop a set of firms names to match with these strings. Our goal is to match our transaction data to all major firms that directly transact with households because we have a relatively complete picture of these firms' revenue.

We start with Compustat and the universe of public firms in a set of industries that meet our criteria of being mostly consumer-facing. These industries include Building materials and garden supply, general merchandise retailers, grocery stores, restaurants, hotels, personal and business services, utilities, home furnishings, apparel, communications, and airlines.[3] In addition, to supplement our set of public firms, we also search the web for lists of large private firms that directly

[3]These correspond to the two-digit SIC codes: 45, 48, 49, 52, 53, 54, 55, 56, 57, 58, 59, 70, 72, and 73. We end up excluding most gasoline stations as their revenue is typically combined with a large refiner or oil producer and thus the consumer-facing business does not provide a good gauge of overall firm revenue or operations.

interface with consumers. For instance, we find lists from sources such as Business Insider, Forbes, and Wikipedia that enumerate the largest firms and retailers in a range of categories, both public and private, and also lists that limit to solely large private firms.[4]

For each of these firms, we then manually search our database of unique transaction strings for transactions that mention the firm name precisely or a range of potential abbreviations and variants of a firm's name. This yields a many to one matching between transaction descriptions and firms. For each firm, there may be many strings that tend to be associated with that firm (e.g. 'wal mart', 'walmart', 'wm super center', 'sams club', 'walmart sacramento', 'walmart joliet', etc.).

Using regular expressions to define our match criteria, our goal is to capture as many true positives as possible while not flagging excessive amounts of false positives. For instance, the term 'subway' will match sandwich purchases at a Subway restaurant but also transactions made at any number of public subway systems around the world or any of the hundreds of small businesses who's name includes the term 'subway'. For this reason, we also often employ limitations in our matching procedure based on retailer category (which is captured in our transaction database) as well as transaction sizes. As one example, when attempting to match Subway sandwich stores, we limit the retailer category of the transaction description to restaurants and the *average* transaction size for the transaction description to under 20 dollars.

Unfortunately, traditional machine learning algorithms are not well suited to the task of mapping these transaction descriptions to firms. Given the huge set of firms in the transaction data (everything from large national retailers to single-establishment stores), automated methods that rely on string-similarity measures tend to produce extremely high rates of false positives. Moreover, many firms' descriptions are not at all similar to their official firm name (e.g. 'tgt' may refer to 'Target Corporation'). For this reason, we mostly rely on manual inspection and experimentation to find descriptions that map to firms. In our entire sample of matched retailers, the mean number of unique text descriptions associated with a given retailer is 176 and the median number is 41.

After working through our sets of large public and private consumer-facing firms, we turn directly to the transaction data to fill in any potential holes in the data. We sort the transaction descriptions by the frequency with which they appear in our data and inspect each of the most

[4]See, for instance the Wikipedia supermarket chains and Wikipedia fast food chains.

frequent 10,000 transaction descriptions. We attempt to map any unmatched transaction descriptions in this set to a firm; generally this firm is one from an industry that we did not previously inspect. For instance, Lyft and Uber appear frequently in our data but are assigned a two-digit SIC industry of 41 (Local And Suburban Transit And Interurban Highway Passenger Transportation). Netflix similarly was not in one of our focused consumer facing industries according to our SIC classification (it is found with two-digit SIC of 78, which mostly contains movie producers).

In the end, we are able to match 558 firms during our sample window. Of these 428 are public and 130 firms are private. While these firms constitute a small fraction of total firms, they are also by far the largest consumer facing firms in the economy. We match our public firm data to Compustat, and rank firms based on their total 2014 revenue. Appendix Table A.4 compares the numerical ranks (with one being the highest), and percentile ranks (with 100% being the highest) of the firms in our matched sample by industry. In all industries, the average firm in our matched dataset is large relative to the average firm in Compustat.

For industries where we have extensive coverage, like airlines, general merchandise, and groceries, we are able to match all 5 of the largest firms. In other industries we have only partial coverage of top firms. For example, we do not match to the Disney Corporation, one of the largest firms in the consumer telecom industry because generally households do not interact directly with the parent company itself (rather they interact through retailers of toys or movie theaters). Similarly, the International Game Technology company is one of the largest entertainment industry firms, but it makes slot machines so has few direct transactions with households. Other firms in our partially covered industries transact mostly with businesses or through webs of subsidiaries that are hard to currently track.

In total, we are able to assign approximately 32% of total spending in our dataset to a particular firm. With additional work, higher numbers of matches could potentially be made with this data.

3 Measuring Customer Bases with Transaction Data

3.1 Customer Characteristics and Revenue

Our matched sample of firms and financial transactions spans a large part of the consumer economy and captures a sizable amount of consumer spending. Table 2 provides some summary statistics

regarding our matched firm-level data. In the first row, we see the the median firm in our sample receives approximately $1.6M from the linked users in our sample in a given quarter. Firm-level spending is skewed towards the largest firms, with the average firm receiving about $8.4M and the largest single firm (Walmart) has observable income from our sample users of approximately $550M per quarter.

The second row in the table displays the fraction of firm's quarterly revenue that we observe among the users in our matched sample. We can only calculate this statistic for public firms with data available on Compustat. On average, we capture about 0.6% of a firm's quarterly revenue (median of 0.4%). There is substantial heterogeneity in the fraction of revenue that we observe in our data – the fraction may be impacted by the portion of a firm's revenue obtained from foreign consumers, whether a firm has substantial business-to-business revenue that is unobserved in our data, and if a firm has a large portion of transactions conducted with cash rather than credit or debit cards. In the third and fourth rows, we note the number of transactions as well as the number of unique users that we can link to a firm in a given quarter. In general, each firm-quarter observation receives tens of thousands of transactions in our data from tens of thousands of users.

3.2 Customer Base Churn and Customer Base Similarity

In this paper, we focus on applications of data regarding two aspects of firm customer bases. The first is the churn in customer base from year to year within a given firm. The second is the similarity (i.e. overlap) of customer bases between two different firms in the same period.

3.2.1 Customer Base Churn

We measure customer base churn as the similarity between the customer base of firm f in year t and the customer base of firm f in year $t-1$, weighted by customer spending at that firm. We define $s_{f,i,t}$ as the share of firm f's revenue in our matched sample that comes from customer i in year t. This definition implies that $s_{f,i,t} \in [0,1]$ and $\sum_i s_{f,i,t} = 1$ for all f and t. We measure churn as:

$$Churn_{f,t-k} = \left(\sum_i |s_{f,i,t} - s_{f,i,t-k}| \right) / (2) \tag{1}$$

where the sum $\sum_i |s_{f,i,t} - s_{f,i,t-k}|$ is taken over all customers that shop at firm f in *either* year t or year $t-k$. In words, churn is the difference in spending shares coming from each customer i between years t and $t-k$. The way it is defined, $\sum_i |s_{f,i,t} - s_{f,i,t-k}|$ can vary between zero and two. A value of zero would imply constant revenue shares, and a constant customer base between years t and $t-k$, while a value of two implies a completely different customer base. We divide this by 2 so churn is normalized to values between 0 and 1. We allow k to vary between 1 and 4 years. Figure 7 plots histograms of this measure across all firms for $k = [1, 4]$. As one would expect, our measure of churn increases over time. That is, the customer base of a firm at time t is more similar to the customer base of that firm at time $t-1$ than at time $t-4$. Over each time horizon, there is substantial spread among firms in how 'sticky' their customer base is. At the most extreme, about 10% of firms see about 90% of their revenue coming from new customers relative to the previous year.

Figure 8 highlights the fact that much of this variation in rates of customer churn over time is driven by systematic differences in rates of churn across industries. Firms in industries like Utilities, Telecom, and Groceries tend to have highly persistent customer base distributions. In contrast, the customers providing revenue in industries such as Hotels, Car Rentals, and Clothing retailers tend to be much less persistent across years. While there exists substantial variation across industries, even in the most homogenous industries, some firms tend to do a much better job at retaining customers and revenue from one year to the next than others. Such variation may be driven by regional concentration, customer loyalty programs, advertising, or other methods.

3.2.2 Customer Base Similarity

The second aspect of a firm's customer base that we focus on is the similarity of firm i's customer base to that of firm j. Again, we define $s_{f,i,t}$ as the share of firm f's revenue in our matched sample that comes from customer i in year t. We define similarity between firms f and j in year t as:

$$Similarity_{(f,j),t} = -\left(\sum_i |s_{f,i,t} - s_{j,i,t}|\right)/(2) + 1 \tag{2}$$

where the sum $\sum_i |s_{f,i,t} - s_{j,i,t}|$ is taken over all customers that shop at *either* firm f or j in year t. As with our churn measure, this sum can vary between zero and two. We multiply by $-1/2$ and add 1 so that a similarity score of one would imply that the firms have the exact same revenue share

from each customer, and a value of zero would imply no overlap in customer bases. We calculate this measure for all firm-firm pairs in our sample at an annual frequency.

Figure 9 displays the average level of customer base similarity within a broad industry group for all firm-firm pairs in that industry. As with the customer base churn metric discussed above, there exists substantial variation in cross-firm similarity across industries. Firms within the Utility industry are the most dissimilar to other Utility firms – which is to be expected as most customers have only a single utility provider and do not vary in their provider much over time. In contrast, restaurants have the highest amount of within-industry cross-firm similarity – over 5 times higher than that of Utility firms. This reflects the fact that many users tend to spend large amounts of money eating out but spread their spending across multiple restaurants rather than focusing on a single restaurant.

Figure 10 focuses on the difference in customer base similarity within an industry versus firm-firm similarity across industries. We note that, on average, within-industry customer base similarity is higher than that across industries (this is seen as the right panel, within-industry firm-firm similarity, is shifted to the right relative to the left panel). That is, many users tend to disproportionately weight their spending towards a particular industry, not simply a particular firm within an industry. However, both panels in this figure have substantial cross-firm variation – for both within- and cross-industry firm-firm pairs we see some that are highly dissimilar and some that are highly similar. Moreover, the set of most similar firms for a given firm tends to span industries.[5]

3.3 Firm Quality and Customer Concentration

Another aspect of firm customer bases that can be easily surmised from transaction-level data is that of the average income of any given consumer-facing firm. Following our work in Baker et al. (2019), we can construct a quarterly index of the average user income of a store's clients, weighted by the amount they spend at that retailer:

$$Quality_{rt} = \frac{\sum_i spending_{irt} * income_{it}}{\sum_i spending_{irt}}$$

[5]For instance, the ten firms with the most similar customer bases to Walmart are: Yum Brands, Dine Brands, Darden Restaurants, Sonic Corp, Netflix, Amazon, Kohls, Dollar Tree, Dominos, and Papa Johns. Among retailers, the ten firms with the most similar customer bases to Walmart are: Amazon, Kohls, Dollar Tree, Bed Bath and Beyond, Autozone, Sally Beauty, Gamestop, Office Depot, Big Lots, and Dicks Sporting Goods.

Where r identifies a retailer, i indexes users, and t refers to a calendar year. Firms in our sample exhibit large differences in this measure, lining up with an ex-ante notion of the firm's quality. Figure 5 shows a selection of customer income distributions for pairs of firms in the same industry. For instance, the bottom right panel displays the distribution of customer income (weighted by spending at the firm) within two grocery stores: Save-a-Lot and Whole Foods. We sort income into $1,000 bins and censor the histogram at $300,000 for visibility. We can see that Whole Foods customers tend to be substantially richer than those of Save-a-Lot, indicating a higher quality firm.

One final illustration of the benefit of linking users to firms using this class of transaction data is the ability to get information not only about levels of spending at a particular firm, but the distribution of spending (i.e. revenue) within a firm across its customers. In Table 3, we display statistics that illustrate how concentrated firm revenue is within its customer base. Looking across broad industry categories, we show that there is a substantial amount of variation in revenue concentration. For instance, the top 5% of customers for a given Utility firm provides approximately 15% of a firm's revenue[6]. In contrast, revenue for hotels and airlines is much more concentrated within their customers, with the highest spending 5% of customers making up almost 30% of their revenue in our sample. This variation in concentration is maintained down the distribution of customers, with the top 20% of customers making up around 40% of revenue in low customer concentration industries and over 75% in high customer concentration industries.

4 Validation

In this section, we provide some evidence that our transaction data provides a meaningful window into the source of firms revenue across several dimensions.

4.1 Compustat Validation

Our first test is to directly compare a firm's official revenue data to the spending that we observe at that firm across all individuals in our sample.

We match all public firms in our sample to their official revenue data obtained from Compustat

[6]Here, we mean the percent of revenue in our matched dataset. In this example, the top 5% of customers make up 15% of the revenue *we can see in our matched dataset*, not 15% of the revenue in Compustat.

across our sample period (2010-2015). This comprises 428 firms (of the 558 in our sample). For each firm-quarter, we observe total spending as well as total reported revenue. Our cleaned sample of individuals contains approximately 1.7 million users, out of a total U.S. population of 320 million (as of 2015). If all firm revenue was obtained directly from consumers located in the United States, we would expect that the spending we observe would make up approximately 0.53% of revenue that these firms report. On average, for firms in our matched sample, we observe an average of 0.6% of quarterly revenue (median of approximately 0.4% of quarterly revenue).

In Figure 2, we plot both logs of the total spending and changes in logged spending derived from these two different sources. While the absolute levels are different owing to the fact that we observe only a sample of total spending, we find a strong correlation between our own spending data and the revenue reported by public firms in relative terms. This holds true in both levels of spending as well as quarter-to-quarter changes within a firm. That is, we do a good job of matching relative sizes of firms as well as the within-firm growth dynamics over time.

Our measure achieves higher rates of correlation and fit when restricting to firms that do not have sizable operations overseas. In addition, we see closer correlations when we exclude firms that have larger fractions of revenue from non-household sources (e.g. if a firm has both business-to-business as well as business-to-consumer divisions).

4.2 Chain Store Guide Validation

We also test whether the geographical distribution of stores and revenue firm-level revenue in our data matches the empirical distribution of their stores. To do this, we utilize data from Chain Store Guide (CSG) database, which tracks the physical locations of retailer branches for a wide range of large regional or national chains. In addition, they include some characteristics about the types of establishments, size of stores, and branch number. We collect CSG data from the entirety of our sample period (2010-2015).

We are able to match 58 firms from the CSG database to the sample of retailers we observe in our transaction data. For each firm, we then construct three values. The first two are derived from our transaction data. First, we simply calculate the fraction of consumer spending that we observe from users in a given state at a particular firm for each year in our sample.

$$FracSpend_{ist} = \frac{\sum_i spending_{irst}}{\sum_i \sum_s spending_{irst}}$$

Where i indexes users, r indexes retailers, s indexes states (and Washington DC), and t represents a calendar year.

Secondly, using the transaction-level description strings, we are able to pick out transactions at particular retailers locations. For instance, a transaction may be labeled as 'McDonalds (Store #391)' rather than simply as 'McDonalds'. We utilize this to construct a measure of the fraction of a retailer's locations in a state each year. We also construct the analog to this variable from the CSG data: the fraction of stores in a given state for a firm-year observation.

We would not necessarily expect a perfect one-to-one relationship between these measures for each retailer. Especially for the fraction of spending we observe, since we do not have establishment level sales data. While a state may have 10% of a retailer's physical stores, those stores may account for 15% of that retailer's national sales. However, on average we would expect a strong relationship between these measures. If we are systematically finding that we under- or over-estimate sales occurring in any particular state, we may be more worried about the representativeness of our sample.

In Figure 3, we display bin-scatter plots of these measures across all state-years in our sample. In the top row, we plot the relationship between the two store level measures (fraction of stores by state-year-retailer in our transaction data against fraction of stores by state-year-retailer in the CSG data). The right panel censors the plot to better highlight the fit among the smaller states. The bottom row displays the relationships between the fraction of spending that we observe for a retailer in a state-year against the fraction of stores from the CSG data in a state-year. Figure 4 breaks down these comparisons by state. In all cases, we see a strong relationship that lies quite close to the 45-degree line, suggesting that we are getting an accurate and unbiased sample of the geographic distribution of spending, on average.

4.3 Firm Quality Validation - Yelp

Lastly, we examine the types of users that patronize a given retailer in our data and compare this to external indicators of retailer quality. We construct a measure of the average income of a firm's customers, as described above. As one validation of both our data and this particular measure,

we then compare this measures of firm quality to data from Yelp.com. From Yelp, we are able to obtain indicators of how expensive the average product at a particular firm is for about two thirds of our sample of firms. For each matched firm, we get a rating between $ and $$$$ that indicates low to high prices, respectively. We regress our measure of firm quality on indicators for these price rankings and report the results in Table 4.

On average, we find that firms that have higher income customer bases in our data tend to be those selling higher priced goods. This is both true overall and in all subcategories of firm that we examine. For instance, relative to the average customer of the lowest priced restaurants ($), the average customer of the highest priced restaurants in our sample ($$$$) tends to have a $24,016 higher annual income.

5 Customer Capital and Firm Outcomes

While churn in firm-level customer bases over time is a new metric with which to assess customer-facing firms, we hope to demonstrate that it is a meaningful indicator, as well. Higher levels of churn in firm customer bases may be a measure of (and potential source for) risk and volatility across firms in similar industries. To test this hypothesis, we run the following regression:

$$Outcome_{i,t} = \alpha + \beta Churn_{i,(t-1,t)} + \text{Ind. FE} + \epsilon_{i,t} \tag{3}$$

where $Churn_{i,(t-1,t)}$ is measured based on each year's customer base relative to the previous year's. To better understand how well our measure of firm churn predicts common firm-level indicators of risk, we examine a range of outcome variables: (1) total volatility, the standard deviation of daily stock returns in that year (2) idiosyncratic volatility, the standard deviation of daily CAPM residuals in that year (3) the beta from a regression of a stock's daily excess returns on the excess returns of the market in a given year and (4) revenue growth, measured as the absolute value of the log change in year-over-year revenue.

Table 5 contains the results. For all the volatility measures, there is a strong positive correlation between the outcome of interest and our measure of churn in the univariate regressions. We then want to evaluate whether our churn measure has marginal explanatory power, over industry fixed effects. The "Ind. FE" specification columns do not include the churn measure, but instead

includes fixed effects for the industry groups: Restaurants, General Merchandise, etc. The "Add Churn" specification keeps the industry-level fixed effects, and adds our churn measure.

In all cases, our churn measure remains statistically significant after including the industry fixed effects. This is a high bar, as we only have 4 years of data for each firm, and firms do not switch industries. Moreover, we find substantial increases in R^2 with the inclusion of churn to a specification with industry-level fixed effects. In the total volatility case, adding the churn measure increases the R^2 by almost 0.1 – an increase of about 40% – relative to just including the industry-level fixed effects. These results suggest that firms which have more churn in their customer bases are riskier and more volatile than other firms in the same industry.

5.1 Churn as Firm-level Indicator for Customer Frictions

In Gourio and Rudanko (2014), the authors examine the role that product market frictions may have in driving firm-level outcomes. They note that firms which 'invest' in a sticky customer base are generating a form of intangible capital. That is, firms can invest up front in acquiring a stable base of customers and then extract value from those customers over time. In a search model of firms where these frictions are present, a firm with a higher degree of customer stickiness can be expected to have higher levels of profits, higher markups, higher market to book value (Q), and higher rates of investment.

To demonstrate this claim empirically, Gourio and Rudanko (2014) utilize the ratio of 'Selling, General, and Administrative' (SG&A) to annual firm sales as a proxy for product and customer frictions. In addition, they utilize the ratio of advertising to sales as an alternative proxy. They conjecture that firms which spend more on SG&A or advertising are doing so to invest in long-run customer acquisition.

With our data, we can directly examine the churn in firm-level customer bases. Rather than relying on a proxy like SG&A spending or advertising, we can directly measure how sticky a particular firm's customer base is from year to year. If SG&A spending represents investment in customer bases to create long-term customers, we would expect to see higher SG&A spending associated with lower levels of firm-level churn. Figure 11 displays the correlation between firm-year SG&A expenses and our measure of firm-level annual customer base churn. Contrary to expectations, across our entire sample, we find a weak positive relationship between these two

variables – firms with higher level of churn tend to spend more on SG&A.

However, this relationship is governed by the composition of our sample, which is highly concentrated on retail firms. If we split our sample into retail and non-retail firms, a clear picture emerges: the relationship between customer churn and SG&A is negative for non-retail firms but highly positive for retail firms. These findings are mirrored when we look at how churn relates to advertising rather than SG&A for the subset of firms that report advertising expenditures.

Gourio and Rudanko (2014) hypothesize that firms generally spend to acquire customers for the long-run, a pattern that seems to play out among non-retail firms. This may in part be driven by the prevalence of long-term contracts (as in cell phone providers) or strong loyalty programs (as in hotels and airlines). However, in the retail industry, a constant battle for market share often drives firms to spend substantial amounts on SG&A and advertising only to win customers from its competitors in the short run. Thus, we may expect that these empirical predictions regarding product market frictions may be more accurately captured by our measure of customer churn than by SG&A within the retail industry.

In Table 6, we turn to the same real firm-level outcomes that were investigated by Gourio and Rudanko (2014) and similarly exclude financial firms and utilities. In columns 1-4, we test whether our measure of customer base churn is related to firm investment, profits, markups, and Q at a firm-year level, including firm and industry fixed effects.[7]. In all cases, we find that the relationship is negative, with higher levels of churn being linked to lower levels for all variables.

In columns 5-7, we turn to testing the relationship between customer base churn and the volatility of investment, profits, and markups. In all cases, the dependent variable is scaled by the variability of Q. Again, we find concurring results: higher levels of churn (lower firm-level customer frictions) are correlated with higher volatility across these metrics.

Table 7 then tests whether our measure of customer churn is associated with differences in dynamics of responses to firm-level shocks. Here, we see how the divergence between SG&A and our measure of churn seen in Figure 11 can affect our results. Columns 1 and 4 essentially replicate the results of Gourio and Rudanko (2014), showing that firms with low levels of SG&A appear to be more like 'classical' no-adjustment-cost firms who respond more strongly to shocks to Q than do firms with higher levels of SG&A spending. Columns 2-3 and 5-6 mimic this specification with

[7]Excluding industry fixed effects, we find similarly sized coefficients

an indicator for a firm being in the top half of firms in terms of customer churn or a continuous measure of churn. We find consistent results – firms with higher levels of churn appear to respond more strongly to changes in Q than low-churn firms.

In columns 7-9, we restrict the analysis to retail firms. Here, the magnitude of the coefficient on the SG&A variable falls by 80% and is statistically indistinguishable from zero. In contrast, the coefficient on the interactions with our measures of customer churn remain statistically significant and of approximately the same magnitude as the results including non-retail firms. In short, our measure of customer churn can more consistently capture firm-level customer search frictions for retail firms.

5.2 Firm-Specific Revenue Declines During COVID-19

We further test this relationship between customer-base churn and firm-specific risk in the recent COVID-19-driven recession. Previous work, such as Baker et al. (2019), has shown that the tendency for households to visit new retailers declines as income declines. This may manifest during a recession as households retrenching into their usual retailers and restaurants and not trying out somewhere they have not visited before. To test this effect, we examine whether firms relying on a steady stream of new customers are more strongly impacted by the recent COVID-19 outbreak and associated policy responses.

During March 2020, city and state governments began unprecedented efforts to halt the spread of COVID-19 by dramatically limiting the ability of retail businesses to remain open and to operate normally. Many businesses were virtually halted or else mandated to operate only remotely. For instance, restaurants were often required to allow only take-out or delivery orders, and many other retail establishments were forced to operate only online, using delivery services or curbside pickup. Moreover, the limits in economic activity sparked a large recession and significant declines in consumer spending.

As our primary transaction data do not extend to 2020, we utilize data from the SafeGraph Data Consortium to examine the impact of these events on consumer spending at a range of retail establishments and how these changes in spending are linked to rates of churn measured at those retailers in earlier years. The SafeGraph data uses data from a range of debit cards to track ag-

gregated levels of daily consumer spending across merchants.[8] We use daily spending data from January 2019 through the end of March 2020 and can observe hundreds of millions of transactions at retailers linked to our previous customer metrics.

In particular, we test whether firms with higher levels of customer churn – as measured by our firm-level metric from 2010-2015 – experienced differential declines in consumer spending as compared to those with low levels of customer churn. Table 8 displays the results of this analysis. Column 1 shows that firms, on average, saw 30% reductions in customer spending during March 2020 as compared to March 2019. Columns 2 and 3 then add in interactions with our measure of customer churn. In general, firms with high levels of customer churn (estimated using the 2010-2015 data) tended to see much larger declines in customer traffic and spending than those with low levels of churn. For example, a firm in the top quartile of churn saw a decline in spending of about 37% while those in the bottom quantile saw declines of only 24%.

These differences are not simply driven by the industry that these firms are in. Using both industry and industry times month fixed effects in columns 4 and 5, we see that the effect persists with similar magnitude. Controlling for the industry-level decline in consumer spending during March 2020, firms in the top quartile of churn tended to see declines in consumer spending of 10 percentage points larger (eg. 35% vs. 25% decline).

This exercise highlights the fact that customer capital, as measured by rates of customer churn, has important implications for both firm behavior and firm performance during the business cycle.

6 Customer Base Overlap and Stock Predictability

6.1 Portfolio Analysis

The connection between firms is still an under-explored area in asset pricing. An exception to this is Cohen and Frazzini (2008), which shows that firms connected via the supply chain have predictable returns. Our measure of customer overlap seems like a natural way to identify economically linked firms. If a set of customers are hit by an economic shock, the collection of firms where these customers shop should be similarly affected. Unlike the supply chain linkages in Co-

[8] In particular, the data is sourced from cards issued by Challenger online banks, payroll cards offered by a range of major employers, and government issued cards (mostly alimony recipients).

hen and Frazzini (2008), which are reported in firms' SEC filings, our measure of customer base overlap is not easily observable. If this information is not fully incorporated into stock prices, it may be possible to form portfolios which generate significant alpha relative to known risk-factors.

To test this, we start with all securities in the CRSP/Compustat merged database, and then restrict to ordinary common shares (sharecodes 10 and 11) traded on major exchanges (exchange codes 1, 2 and 3). We also remove financial firms (SIC codes 6000-6999) and utilities (SIC codes 4900-4999). After matching this subset to our customer-base overlap data, we have about 250 firms per month between 2010 and 2018. We form five portfolios each month using the following procedure. First, we compute the average overlap between firms' customer bases for each pair of firms in our sample. We compute this average using the average of annual overlap between 2011-2014, as these are the only years in our sample with four quarters of data. We use a single average, even though this introduces a look-ahead bias in our portfolio formation, as the overlap does not change much over time.

Each month, we identify the 10 firms with the highest overlap for each firm in the matched dataset. We then form a value-weighted portfolio of these 10 firms, and calculate the return of this portfolio over the past quarter. We then sort firms into 5 portfolios: Portfolio 1 (low) has firms whose 10 most overlapping firms had the lowest stock returns over the past quarter. Portfolio 5 (high) has firms whose 10 most overlapping firms had the highest stock returns over the past quarter. We then form a hedge portfolio which is long portfolio 5 and short portfolio 1. We want to test whether the return of firms with high customer-base overlap has predictive power for future returns, adjusting for known risk-factors. We regress the returns of our portfolios on the 5 Fama-French factors (Fama and French (2015)) and a momentum factor (see e.g. Jegadeesh and Titman (1993)) obtained from Ken French's website.

Table 9 contains the results. Alpha is monotonically increasing from the Low to High portfolios. Further, our hedged portfolio has a large and statistically significant alpha of almost 1% per month. This suggests that when firms with similar customer bases to a given firm j have high (low) returns, firm j will likely have high (low) returns in the future. At this point, it is not clear whether this is alpha a risk-premium or an anomaly. To our knowledge, there is no theoretical model of asset prices with heterogeneous/overlapping customer bases, but we conjecture the effect we find is an *anomaly*. Given that our data is not publicly available, it would not be surprising if

this information was not fully incorporated into stock prices.

As mentioned above, our portfolio formation process involves some look-ahead bias. We compute the overlap in customer bases one time using all the data between 2011 and 2014, and apply that to portfolio formation between 2010-2018. Table A.1 forms portfolios, but without a look ahead bias. We use the overlap in year t to form portfolios in year $t+1$. For example, we use overlap data from 2011 to form portfolios in 2012. This shrinks our sample, as we do not extend portfolio formation back to 2010, or extend forward to 2016-2018. Even in this smaller sample, and without the look-ahead bias, the alphas are monotonically increasing from the low to high portfolios. Further, the alpha on the hedge portfolio is almost unchanged in magnitude, and is still statistically significant. This suggests that this look-ahead bias is not driving our results.

Another concern is that our measure of customer overlap is picking up a firm characteristic already known to predict returns or risk premia. An obvious one is momentum, as it's possible that the returns of similar firms are highly correlated with a firm's own past returns. This is unlikely to drive our results, however, as we are already controlling for the momentum factor in all the asset pricing regressions.[9]

Another possible proxy for customer base overlap is correlation of stock returns. To test this, we compute the correlation of each pair of firms' daily stock returns from 2011-2014. Then, each month, we identify the 10 most correlated firms. We then repeat the procedure for forming 5 portfolios from above, except we use the 10 most correlated firms instead of the 10 firms with the highest overlap on customer base. Portfolio 1 (low) has firms whose 10 most correlated firms had the lowest returns over the past quarter. Portfolio 5 (high) has firms whose 10 most correlated firms had the highest returns over the past quarter. Table A.2 contains the results. There is no pattern in the alphas from low to high. This suggests that our results are picking up something independent of correlation.

Despite the results in Table A.2, it's possible that our results are still related to past correlation

[9]In unreported results, we perform a 2-by-2 double-sort on own firm returns from $t-12$ to $t-2$ as in Jegadeesh and Titman (1993), and returns of firms with high customer base overlap over the past quarter. We find that the returns on portfolios that go long firms with overlapping firms which have high returns, and short firms with overlapping firms which have low returns has a positive alpha regardless of whether we restrict to only low past-return/momentum firms, or high past-return/momentum firms. This is not surprising, given the poor performance of momentum strategies between 2010 and 2018.

in stock returns. To further rule out this channel, we perform a double sort. The first sort is on performance of high customer base similarity firms with above/below median past returns. The second sort is on performance of high past stock market correlation with above/below median past returns. We then form two hedge portfolios on the overlap dimension. Table A.3 contains the results. Both hedge portfolios have statistically significant alphas, again suggesting that our results are not driven by correlation in stock returns among firms with high customer base overlap.

6.2 Earnings Announcements

To understand the mechanism behind the results in Table 9, we examine days where we know fundamental information about firms is released: earnings announcements.

To make things easier to understand, in this subsection, we explain everything from the perspective of a single example firm, Wal-Mart (WMT). All the regressions, however, use data from all the firms in our dataset that we can match to IBES. We require matching to IBES for two reasons: (1) IBES provides the *time* of each earnings announcement. This is important, because it lets us determine the first day that investors could trade on that information during normal hours – we call this the effective earnings announcement date. For example, if earnings were released at 8AM on a Monday, we would identify that as the effective earnings date. If earnings were released at 5PM on a Monday, the next trading day would be the effective earnings date. Table 10 contains the results. In all the tests that follow, we restrict to firms which have the same fiscal period end as WMT (although not necessarily the same fiscal year end), and that release earnings in the same quarter as WMT.[10]

We use a definition of standardized unexpected earnings (SUE) as the year-over-year (YOY) earnings growth divided by the standard deviation of YOY earnings growth over the previous 8 quarters (see e.g. Novy-Marx (2012)). We are interested in whether earnings growth in firms with high customer base overlap with WMT has predictive power for earnings growth at WMT. Column 1 is a regression of WMT's SUE on the SUE of the 20 firms with the highest overlap to WMT, which released earnings before WMT in a given calendar quarter. Column 2 is a regression of the SUE of the 20 firms with the highest overlap to WMT on WMT's SUE, but which released

[10]This essentially excludes firms which release earnings late. A firm releasing news late is news in and of itself, see e.g. Begley and Fischer (1998).

earnings after WMT in a given calendar quarter. Column 1 implies that when firms with similar customers to WMT have high earnings growth, and report earnings before WMT, WMT also has high earnings growth. Column 2 says that when WMT has high earnings growth, high overlap firms which report later in the quarter also have high earnings growth.

Having shown predictability in fundamentals, we want to show predictability in stock returns around earnings announcements. Define earnings-day returns as the cumulative market-adjusted log returns from $t-5$ to $t+1$ where t is an earnings announcement date. We define market-adjusted returns as in Campbell et al. (2001): The difference between the excess return on the stock, and the return on the market factor from Ken French's data library. We are interested in whether high earnings day returns for firms with high customer base overlap with WMT has predictive power for earnings day returns for WMT.

Column 3 is a regression of WMT's earnings day returns the earnings day returns of the 20 firms with the highest overlap to WMT, which released earnings before WMT in a given calendar quarter. Column 4 is a regression of the earnings day returns of the 20 firms with the highest overlap to WMT on WMT's earnings day returns, but which released earnings after WMT in a given calendar quarter. Column 3 implies that when firms with similar customers to WMT have high earnings day returns, and report earnings before WMT, WMT also has high earnings day returns. Column 4 says that when WMT has high earnings day returns, high overlap firms which report later in the quarter also have high earnings day returns.

Finally, we are interested in how analysts covering WMT, and firms with overlapping customer bases, react to the release of new information. Define forecast (in)accuracy as the absolute difference between actual earnings per share and the average analyst forecast of earnings per share, normalized by the share price at the time of the earnings announcement. We are interested in whether analyst accuracy for firms with high customer base overlap with WMT has predictive power for analyst accuracy for WMT. The logic is that analysts could use large surprises at firms with large overlap to correct their forecasts for WMT. If this were true, when those other firms had a large surprise, relative to analyst estimates, we would expect WMT to have a smaller surprise.

Column 5 is a regression of WMT's analyst accuracy on the analyst accuracy of the 20 firms with the highest overlap to WMT, which released earnings before WMT in a given calendar quarter. Column 6 is a regression of the analyst accuracy of the 20 firms with the highest overlap to WMT

on WMT's analyst accuracy, but which released earnings after WMT in a given calendar quarter. Both columns are insignificant, which suggests that analysts do not use this overlap information to update their forecasts.

7 Conclusion

Using credit and debit card transaction data, this paper demonstrates that it is possible to construct accurate pictures of firm revenue, growth, geographic dispersion, and customer base characteristics at a highly granular level. We use this data to develop two new measures of firm customer bases. First, we generate an index of the rate of churn in a firm's customer base over time. Second, we construct a metric of the pairwise similarity between two firms' customer bases.

We show that these measures provide important insight into the behavior of both real firm decisions and firm asset prices. Rates of customer churn affect the level and volatility of firm-level investment, markups, and profits and affect how quickly firms respond to shocks. In addition, similarity between firms' customer bases highlights one under-explored type of correlated shocks – we demonstrate that significant alpha can be generated using a trading strategy that exploits our index of firm-firm customer base similarity. We note that these measures are possible to construct by researchers using an increasingly accessible class of financial transaction data and encourage researchers outside fields like household finance and macroeconomics to leverage transaction data in order to answer questions regarding consumer-facing firms.

References

Sumit Agarwal and Wenlan Qian. Consumption and debt response to unanticipated income shocks: Evidence from a natural experiment in singapore. *American Economic Review*, 104(12):4205–4230, 2014.

Sumit Agarwal, Wenlan Qian, and Xin Zou. Disaggregated sales and stock returns. *Working Paper*, 2020.

Deniz Aydin. Consumption response to credit expansions: Evidence from experimental assignment of 45,307 credit lines. *Working Paper*, 2019.

Pierre Bachas, Paul Gertler, Sean Higgins, and Enrique Seira. How debit cards enable the poor to save more. *Working Paper*, 2019.

Scott Baker, Lorenz Kueng, Steffen Meyer, and Michaela Pagel. Measurement error in imputed consumption. *Working Paper*, 2020.

Scott R. Baker. Debt and the Response to Household Income Shocks: Validation and Application of Linked Financial Account Data. *Journal of Political Economy*, 126(4):1504–1557, 2018.

Scott R. Baker, Brian Baugh, and Lorenz Kueng. Income Fluctuations and Firm Choice. *Working Paper*, 2019.

Brian Baugh, Itzhak Ben-David, and Hoonsuk Park. Can taxes shape an industry? evidence from the implementation of the âĂIJamazon taxâĂİ. *Journal of Finance*, 73(4):1819–1855, 2018.

Brian Baugh, Itzhak Ben-David, Hoonsuk Park, and Jonathan Parker. Assymetric consumption smoothing. *Working Paper*, 2020.

Joy Begley and Paul E Fischer. Is there information in an earnings announcement delay? *Review of accounting studies*, 3(4):347–363, 1998.

John Y Campbell, Martin Lettau, Burton G Malkiel, and Yexiao Xu. Have individual stocks become more volatile? an empirical exploration of idiosyncratic risk. *The Journal of Finance*, 56(1):1–43, 2001.

Lawrence J Christiano, Martin Eichenbaum, and Charles L Evans. Nominal rigidities and the dynamic effects of a shock to monetary policy. *Journal of political Economy*, 113(1):1–45, 2005.

Lauren Cohen and Andrea Frazzini. Economic Links and Predictable Returns. *Journal of Finance*, 63, 2008.

Janice Eberly, Sergio Rebelo, and Nicolas Vincent. What explains the lagged-investment effect? *Journal of Monetary Economics*, 59(4):370–380, 2012.

Eugene F Fama and Kenneth R French. A five-factor asset pricing model. *Journal of financial economics*, 116(1):1–22, 2015.

Peter Ganong and Pascal Noel. Consumer spending during unemployment: Positive and normative implications. *American Economic Review*, 109(7):2383–2424, 2019.

Francois Gourio and Leena Rudanko. Customer Capital. *Review of Economics Studies*, 81, 2014.

Gerard Hoberg and Gordon Phillips. Product market synergies and competition in mergers and acquisitions: A text-based analysis. *Review of Financial Studies*, 23(10):3773–3811, 2010.

Narasimhan Jegadeesh and Sheridan Titman. Returns to buying winners and selling losers: Implications for stock market efficiency. *The Journal of finance*, 48(1):65–91, 1993.

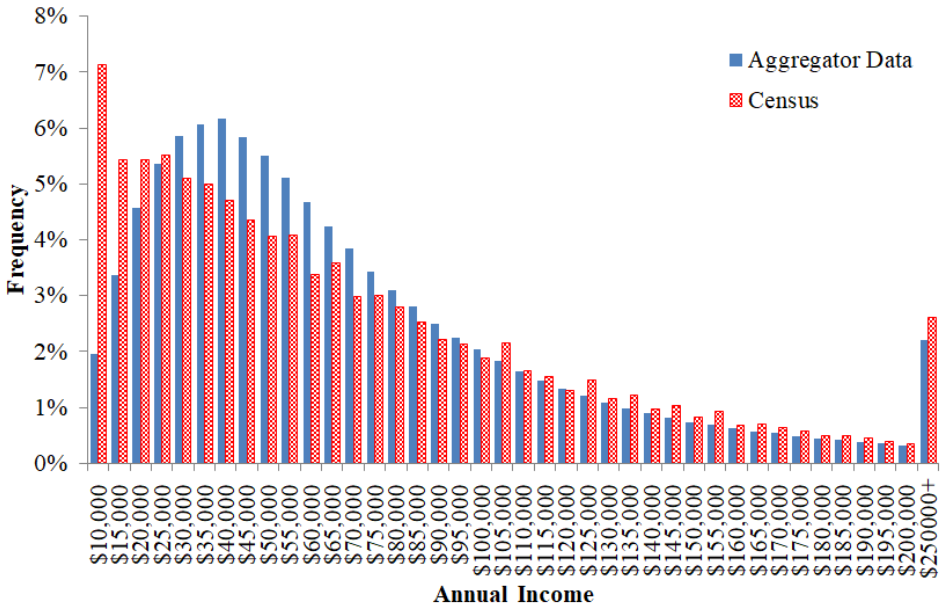
Lorenz Kueng. Excess Sensitivity of High-Income Consumers. *Quarterly Journal of Economics*, 133(4):1693–1751, 2018.

Paolina C. Medina. Selective attention in consumer finance: Evidence from a randomized intervention in the credit card market. *Working Paper*, 2020.

Robert Novy-Marx. Is momentum really momentum? *Journal of Financial Economics*, 103(3): 429–453, 2012.

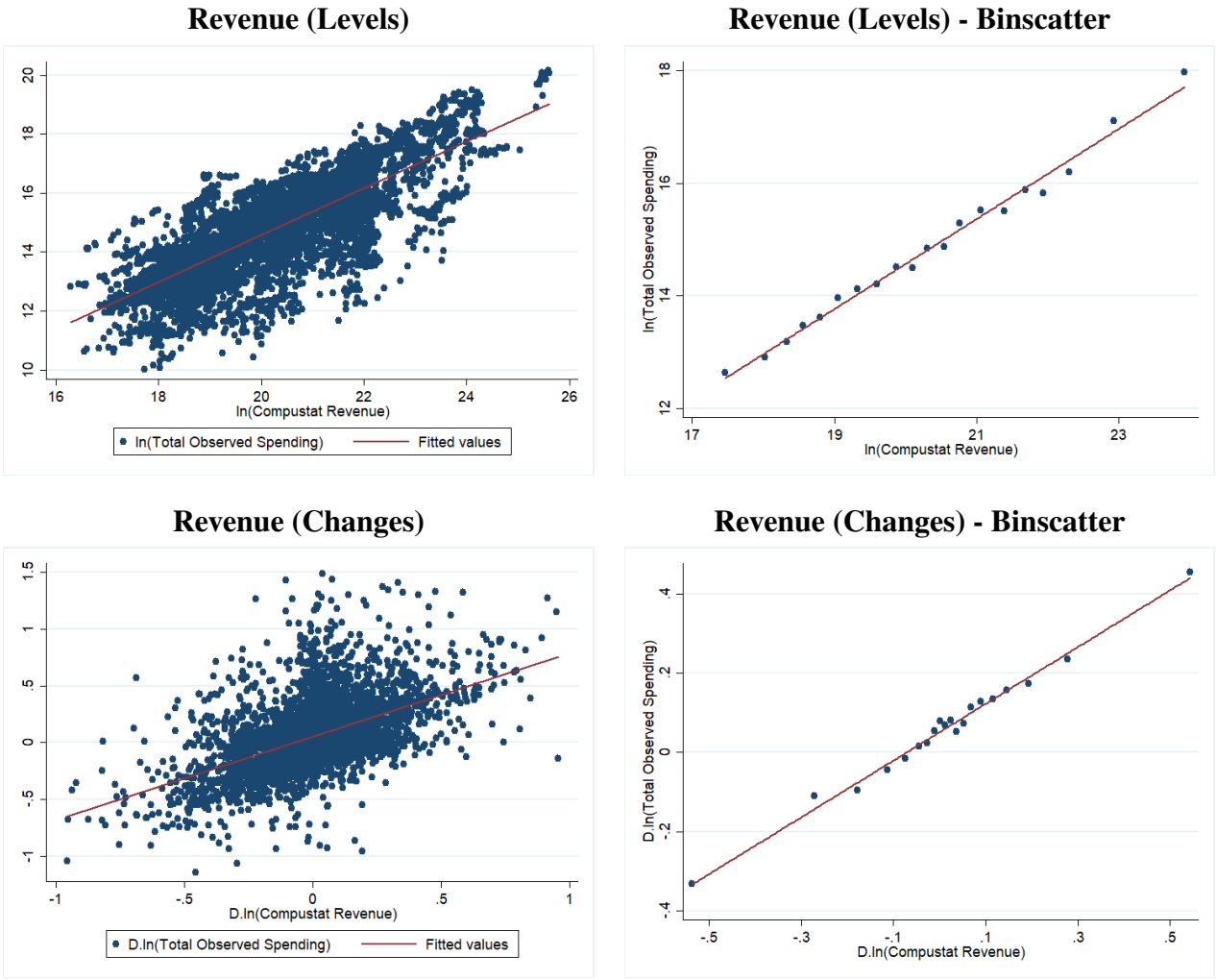
Arna Olafsson and Michaela Pagel. The liquid hand-to-mouth: Evidence from personal finance management software. *Review of Financial Studies*, 31(11):4398–4446, 2018.

Figure 1: Income Distribution - Aggregator Data vs. U.S. Census



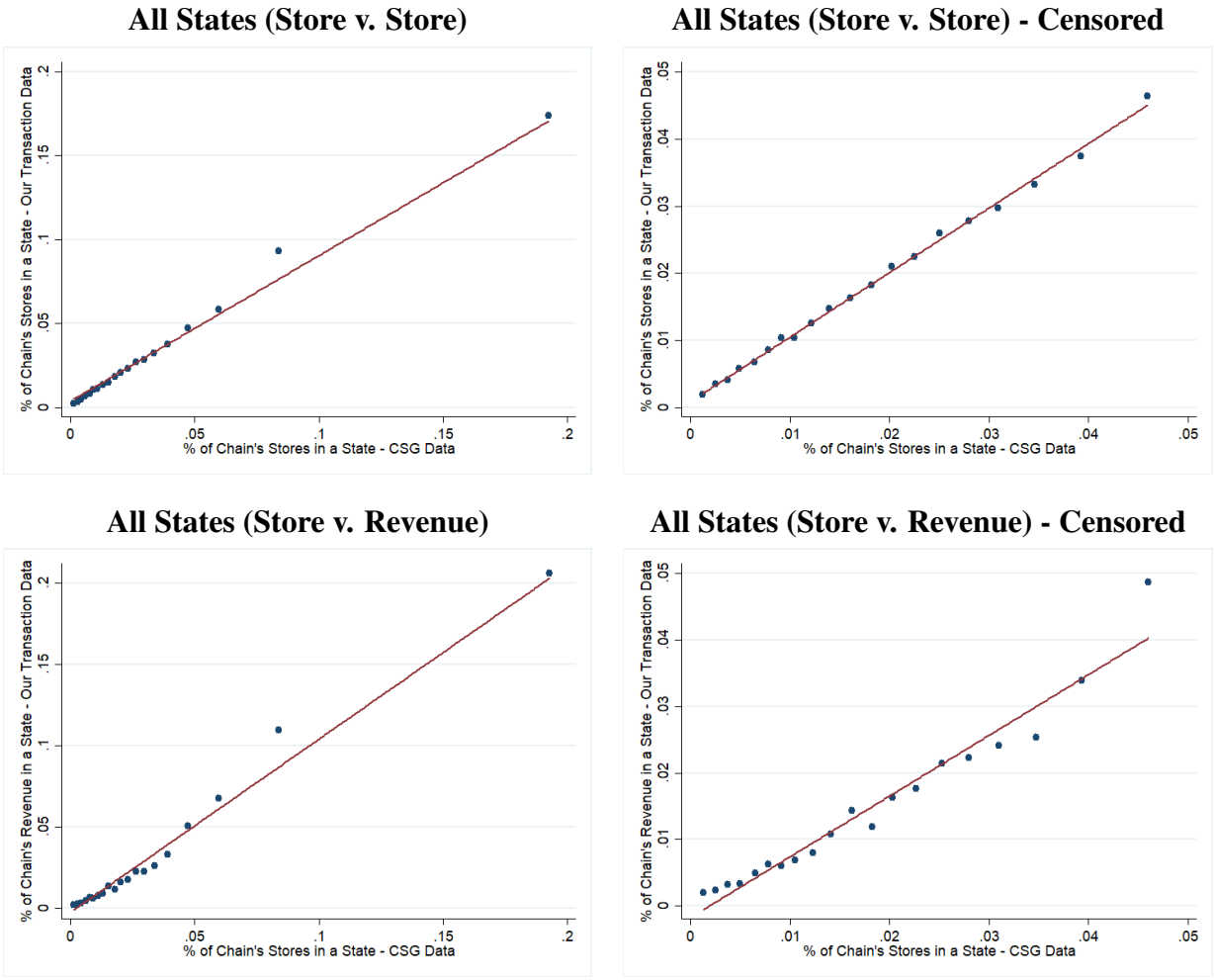
Notes: This figure compares the distribution of 2014 income of the account aggregator and the U.S. Census. The Census data uses the variable *HINC-06* and is available for download at census.gov. The difference in distributions at the bottom end of the income distribution is due to censoring of zero income users in our dataset. See Section 2 for more details.

Figure 2: Comparison Between Reported Revenue and Observed Spending



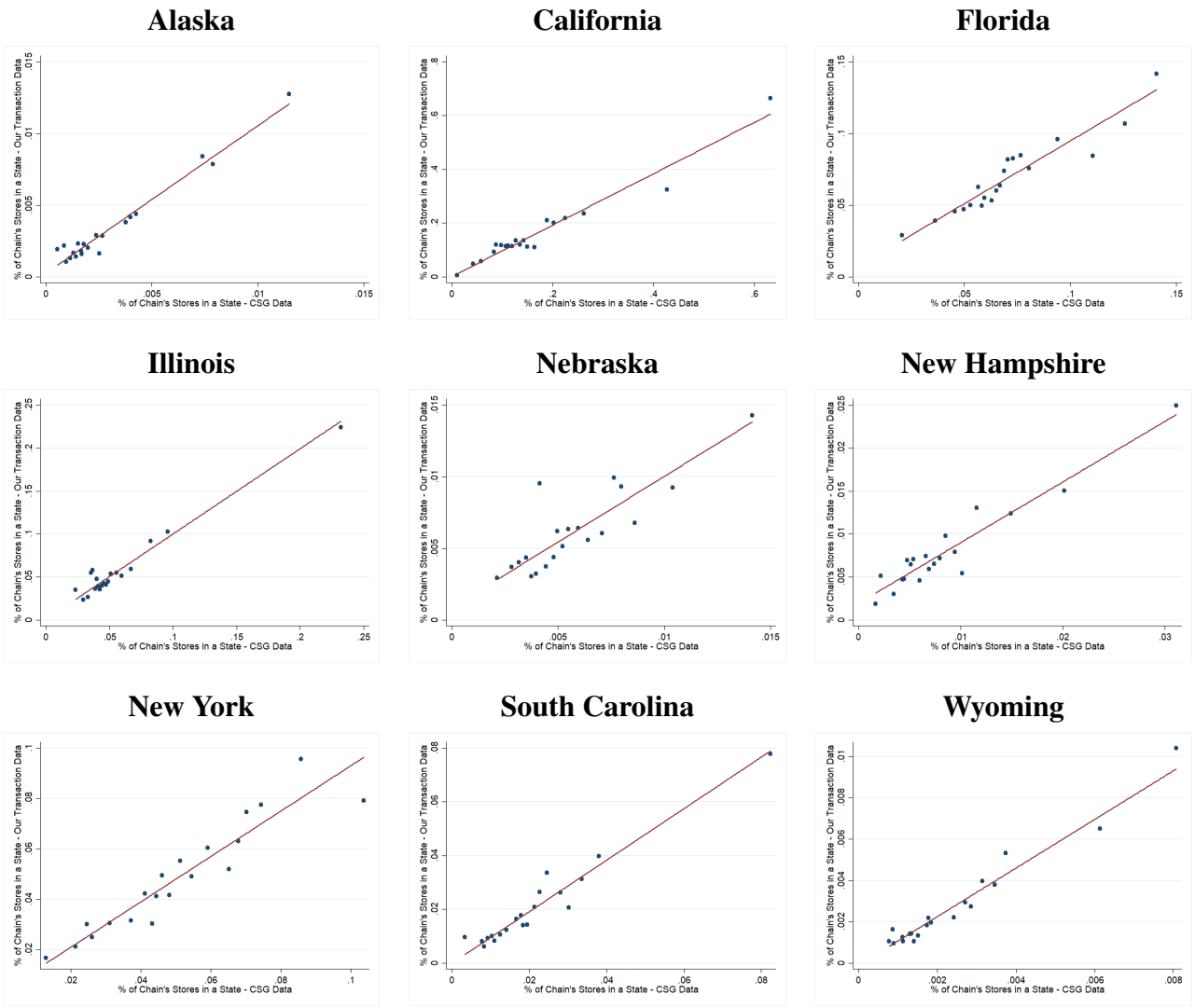
Notes: These graphs show the relationship between firm-level revenue measured in two ways: through Compustat and as observed in our transaction data. Each dot denotes a firm-quarter observation. Along the x-axis, we measure $ln(Revenue_{it})$ obtained from Compustat. Along the y-axis, we measure the total spending observed at a firm in a quarter within our transaction database. The top two panels examine levels of revenue and observed transaction spending. The bottom two panels examine changes in revenue and observed transaction spending.

Figure 3: Geographic Concentration - Transaction Revenue Data and Chain Store Guide Data



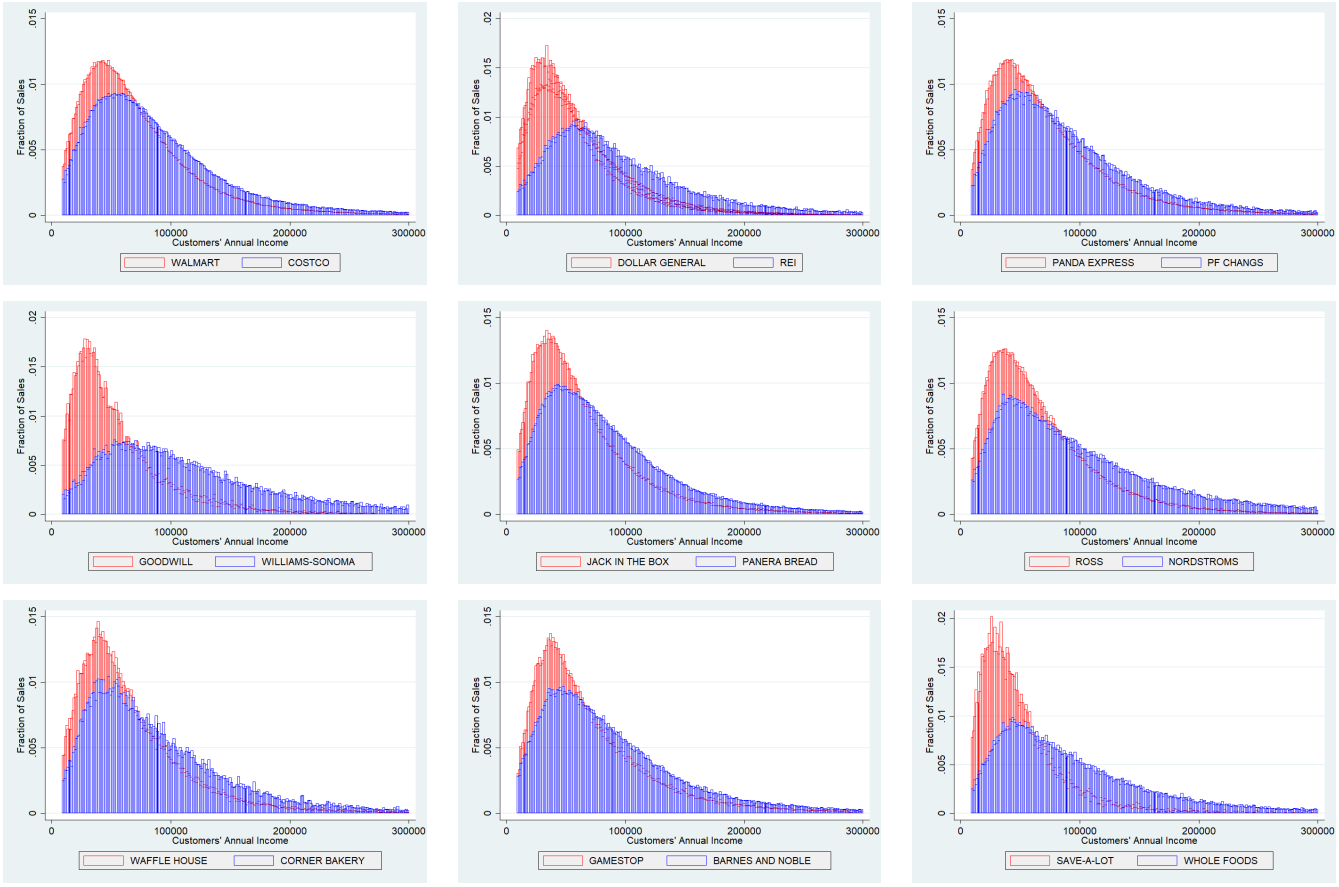
Notes: The graphs demonstrate the relationship between geographic concentration within a firm in two different ways. The first, measured on the x-axis, uses data from Chain Store Guide data and limits our sample primarily to retail firms. The x-axis measures the fraction of a firm's stores that are in a given state in a year (an observation is a firm-state-year). The y-axis measure uses data from our transaction data base and measure the fraction of spending at a retailer that is conducted by users living in a given state. Data covers all retailers able to be matched between samples and spans all 50 states, 2011-2014.

Figure 4: Geographic Concentration - Transaction Store Data and Chain Store Guide Data, Selected States



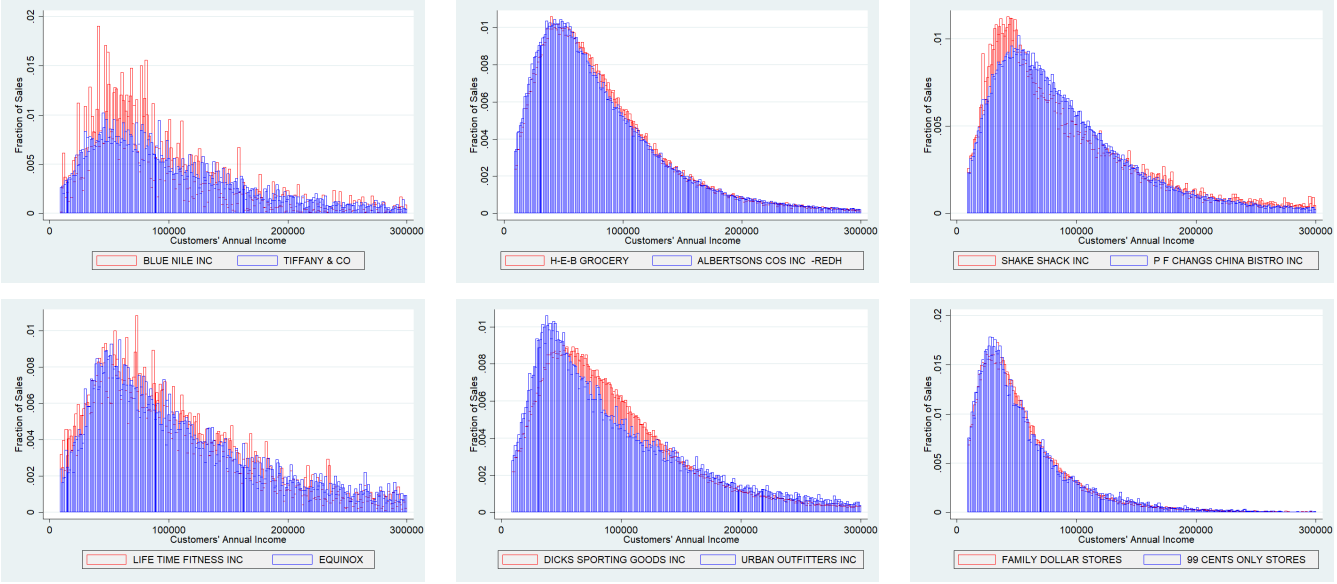
Notes: The graphs demonstrate the relationship between geographic concentration within a firm in two different ways. The first, measured on the x-axis, uses data from Chain Store Guide data and limits our sample primarily to retail firms. The x-axis measures the fraction of a firm's stores that are in a given state in a year (an observation is a firm-state-year). The y-axis measure uses data from our transaction data base and measure the fraction of spending at a retailer that is conducted by users living in a given state. For each graph, the data spans all retailers operating in the listed state in our matched sample, 2011-2014.

Figure 5: Income Distribution of Customerbase, Firm-level Comparisons



Notes: Figures demonstrate the distribution of income among customers for a selected sample of firms. Customer's are dollar-weighted by sales at a firm, so a user spending $500 at a firm will have double the weight in the histogram as a user spending $250. Annual income is binned in $1,000 increments and is censored at $300,000 for illustrative purposes. In each panel, two firms of similar types are compared. Data spans 2010-2015.

Figure 6: Income Distribution of Customerbase, Firm-level Comparisons (continued)



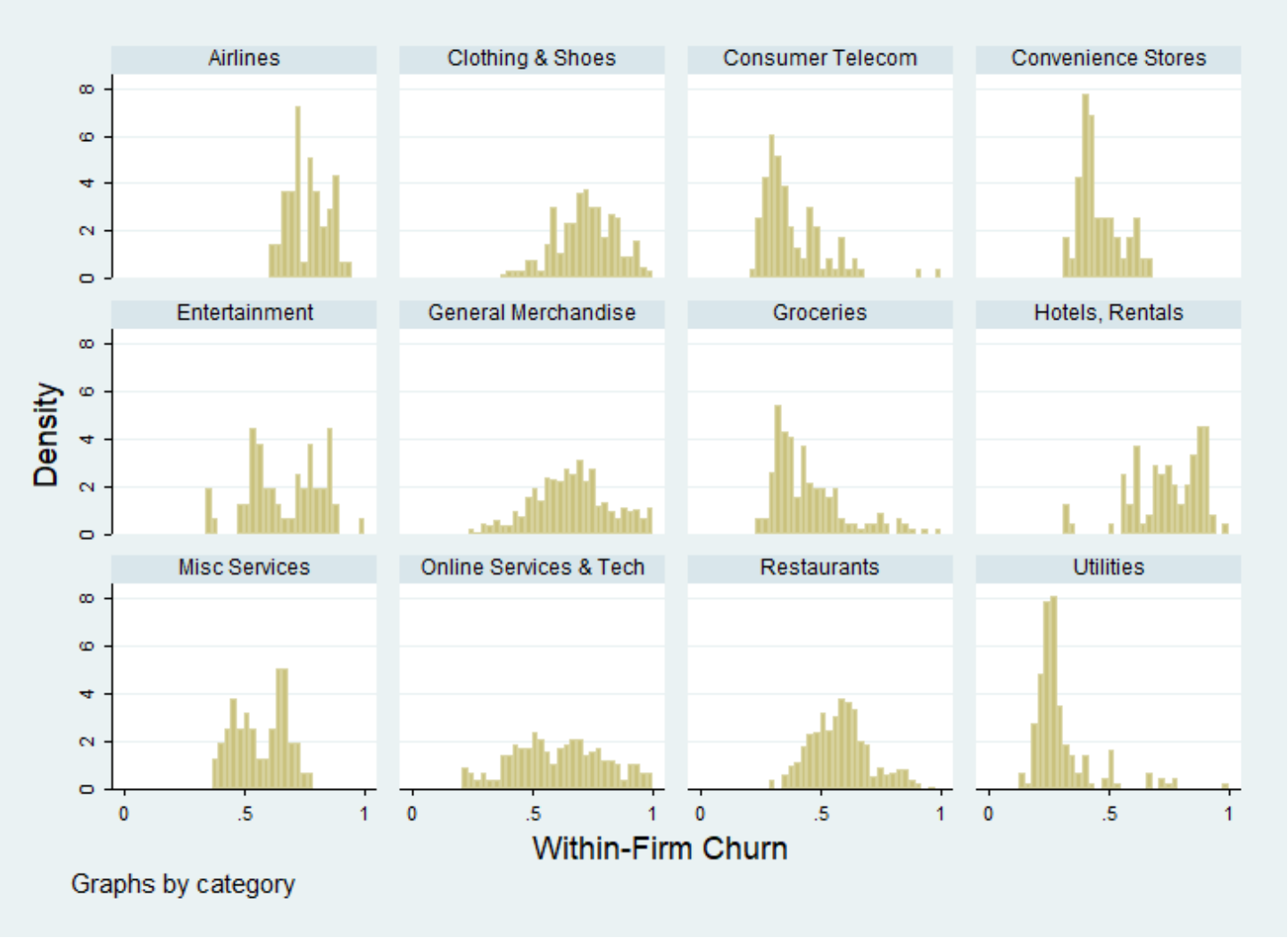
Notes: Figures demonstrate the distribution of income among customers for a selected sample of firms. Customer's are dollar-weighted by sales at a firm, so a user spending $500 at a firm will have double the weight in the histogram as a user spending $250. Annual income is binned in $1,000 increments and is censored at $300,000 for illustrative purposes. In each panel, two firms of similar types are compared. Data spans 2010-2015.

Figure 7: Customer-Base Similarity Within Firm Over Time



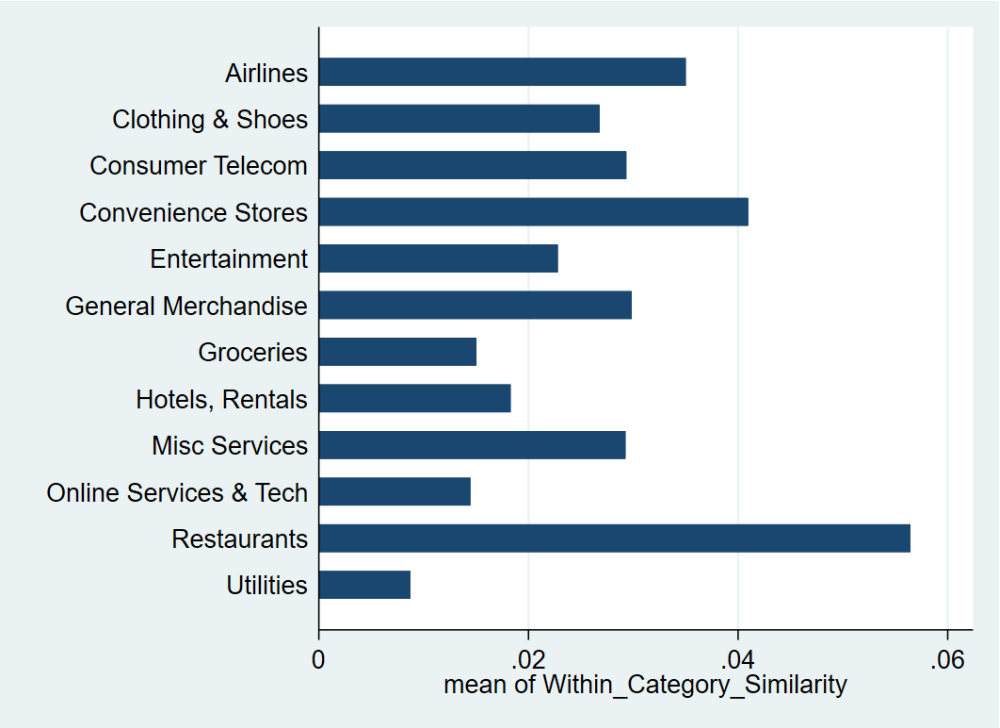
Notes: Each panel denotes the distribution of customer base churn over time across all firms in our sample. Churn is measured as the dollar-weighted overlap between the customer base of a firm f in year t and the customer base of firm f in year $t-x$ where x is between 1 and 4 and is labeled above each panel. Overlap is scaled between 0 and 1 where 1 is an identical customer base and 0 is no overlap between customer bases across years.

Figure 8: Customer-Base Annual Churn, By Industries



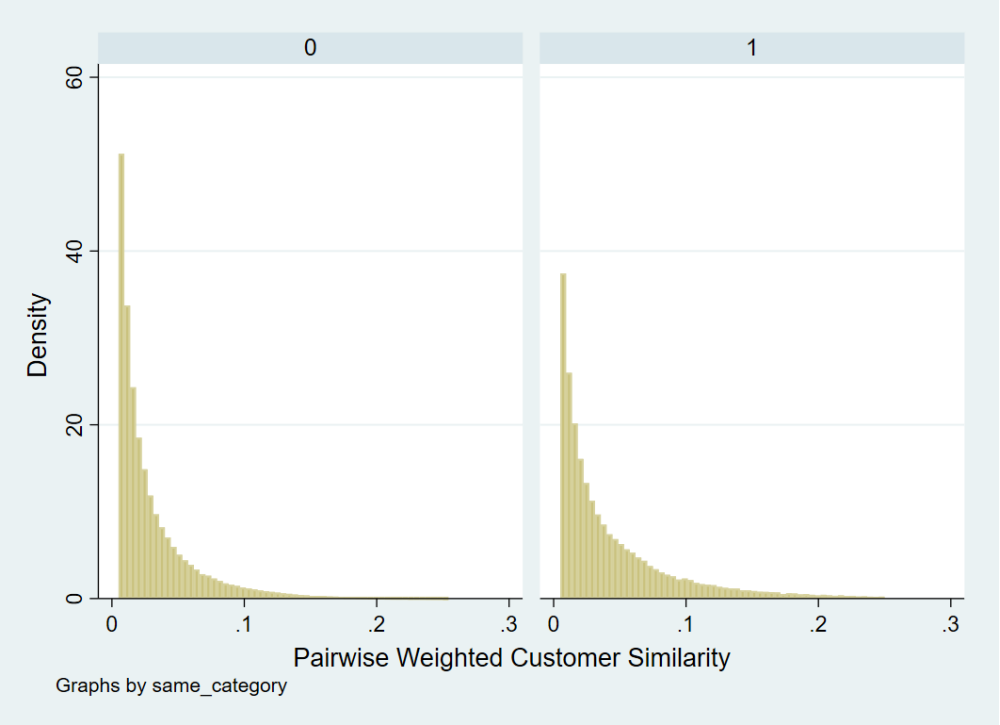
Notes: Each panel denotes the distribution of customer base churn over time across all firms in a given industry grouping in our sample. In this figure, churn is measured as the dollar-weighted overlap between the customer base of a firm f in year t and the customer base of firm f in year $t-1$. Overlap is scaled between 0 and 1 where 1 is an identical customer base and 0 is no overlap between customer bases across years.

Figure 9: Similarity of Firm Customer Bases Within Category, by Category



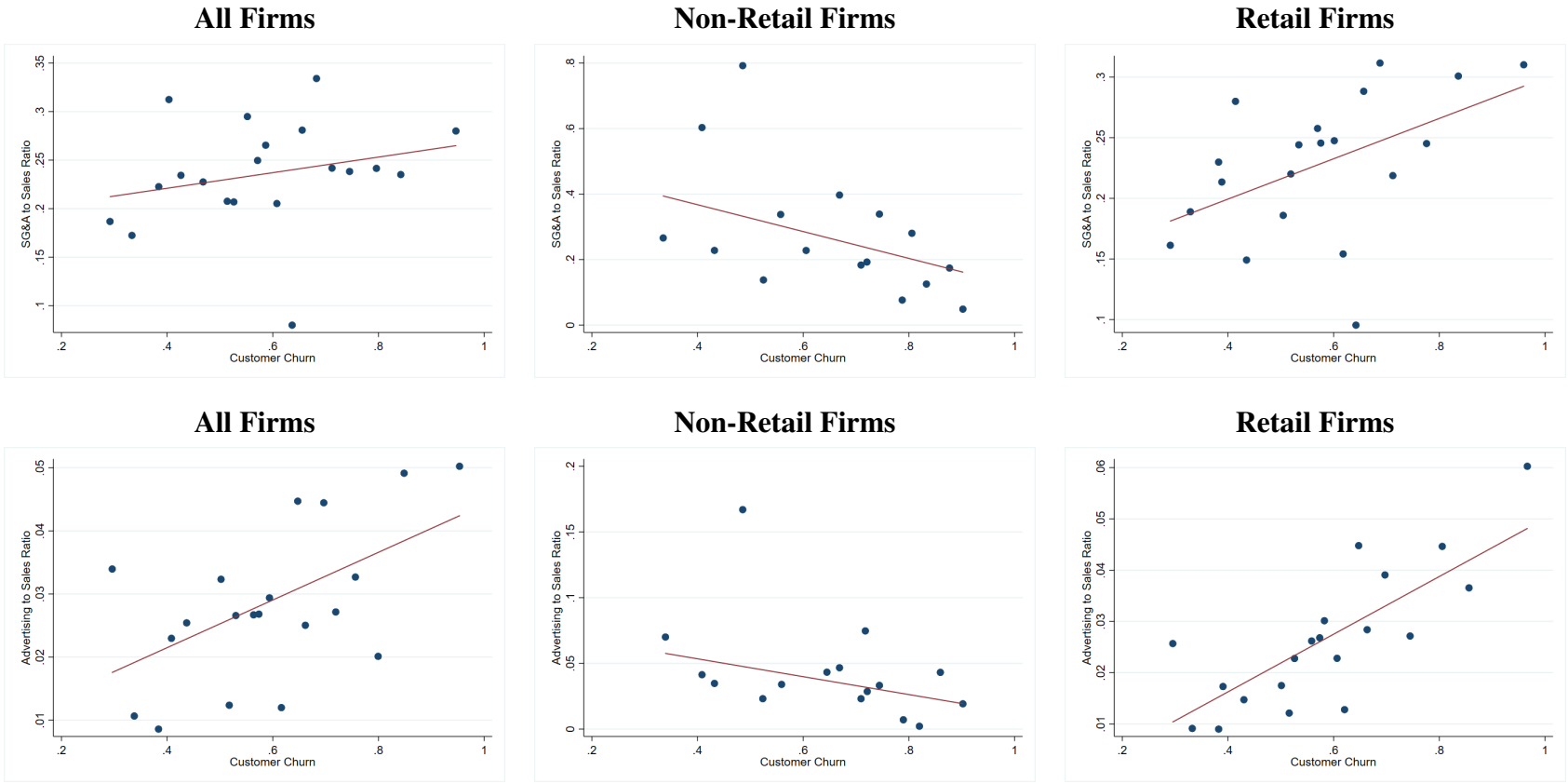
Notes: Bars denote the average cross-firm similarity within the listed industries. That is, the similarity between firm i and firm j who are both operating in broad industry classification x.

Figure 10: Similarity of Firm Customer Bases Within Category and Outside of Category



Notes: The left panel displays a histogram that describes the distribution of firm-firm similarity over all annual firm-firm pairs where firm i and firm j are in different broad industry groups. The right panel displays a histogram that describes the distribution of firm-firm similarity over all annual firm-firm pairs where firm i and firm j are in the same broad industry.

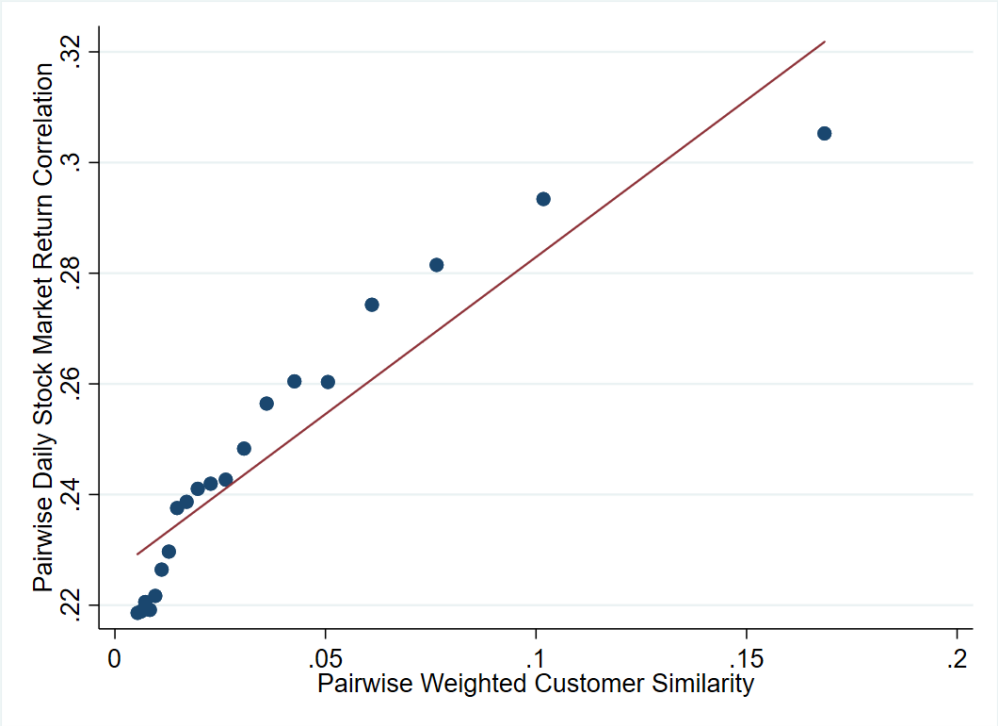
Figure 11: S,G&A Expenses, Advertising Expenses, and Customer Churn



Notes: Retail firms defined as public firms in our sample with a one-digit SIC code of '5'. SG&A expenses and Avertising expenses obtained for all firms with non-missing data in Compustat. Customer churn scaled between zero and one and is measured as the similarity of a firm's customer base at time t relative to the customer base at time $t-1$, weighted by customer spending. Observations in the underlying data are firm-year. Plotted data cover 2011-2014 to exclude partial-year observations.

Figure 12: Stock Correlation and Similarity of Firm Customer Base

Average Annual Correlations and Similarity



Notes: Plot represents a bin-scatter of firm-firm pair annual observations. Stock correlation is measured as the average pairwise daily correlation in stock returns for a given year in our sample period. Similarity in customer base is measured as the fraction of spending at firm i that overlaps with with spending at firm j (e.g. is sourced from the same customers) and is averaged across the entirety of our sample period. Firm customer base similarity is scaled to between zero and one.

Table 1: Examples of Transaction String Data

Description	Count of Txns	Average Txn Amount	Frac Debit	Avg Loose Recurring
home depot	11,002,662	74.31	0.911	0.001
starbucks corpx	8,676,113	7.14	0.999	0.007
jack in the box	3,035,066	8.91	1.000	0.005
aeropostale	327,696	41.53	0.948	0.001
duane reade th avenew	160,318	18.72	1.000	0.004
bos taxi med long island cny	46,648	17.68	1.000	0.002
sbc phone bill ca bill payment	22,248	83.07	1.000	0.132
golden pond brewing	2,385	38.98	1.000	0.001
cross bay bagel	1,542	15.46	1.000	0.000
lebanese taverna bethe	1,542	68.44	0.999	0.005
racetrac purchase racetrac port charlot	1,357	31.32	1.000	0.007
trader joes rch palos vr	1,273	41.91	1.000	0.000
chevys fresh mex aronde	956	36.83	1.000	0.000
gracey s liquor	113	15.99	1.000	0.018

Notes: Table denotes sample transaction descriptions from our database of financial transactions. Each panel displays the cleaned description string (e.g. removing numerics), the number of observations of that string in our data, the average transaction amount for that description string, the fraction of transactions that are debited from an account (instead of credited), and the fraction of transactions that are similar to a previous transaction to that description within a user.

Table 2: Summary Statistics, by Firm-Quarter

Variable	# Obs.	Mean	10%	25%	50%	75%	90%
Observed Spending	10,528	$8,368,492	$51,955	$439,811	$1,616,576	$5,324,263	$16,539,201
$\frac{ObservedSpending}{CompustatRevenue}$	6,751	0.0061	0.0002	0.0013	0.0041	0.0076	0.0127
Number of Transactions	10,528	204,425	734	6,964	39,472	131,970	423,665
Unique Users	10,528	66,317	353	4,082	19,969	64,603	171,473

Notes: Table reports basic summary statistics regarding the 558 matched firms in our sample. Compustat revenue data only available for the subset of public firms in our sample. An observation is a firm-quarter. Quarters with no observed transactions for a given firm are dropped.

Table 3: Customer Base Concentration, by Industry

Category	# Obs.	HHI	Top 5% Share	Top 10% Share	Top 20% Share
Clothing & Shoes	207	0.57	24.8%	37.7%	55.1%
Consumer Telecom	59	0.62	17.9%	30.3%	49.3%
Convenience Stores	44	0.70	40.6%	56.5%	73.2%
Entertainment	56	1.50	25.2%	37.7%	55.1%
General Merchandise	462	0.81	29.1%	43.1%	61.2%
Groceries	166	1.51	42.8%	59.9%	77.3%
Hotels, Rentals, Airlines	96	1.16	29.2%	42.5%	60.7%
Misc Services	59	0.57	24.8%	37.7%	55.8%
Online Services & Tech	126	1.12	24.7%	36.9%	53.9%
Restaurants	369	0.38	27.9%	41.1%	57.9%
Utilities	116	0.83	15.5%	26.7%	44.6%

Notes: Table reports summary statistics across firms in a range of industry groupings. An observation is a firm-year. HHI is within-firm concentration in customer dollars. HHI is measured as the sum of squared fractions of revenue obtained from each customer, multiplied by 10,000. In this table, we equally weight firm-years but remove firms with fewer than 7,500 observed customers in a year.

Table 4: Firm Quality Index and Yelp Ratings

	(1)	(2)	(3)	(4)	(5)
VARIABLES	All Stores	Restaurants	General Stores	Clothing	Groceries
Yelp - $$	11,845***	8,176***	11,364***	18,135***	8,240***
	(402.7)	(622.9)	(833.4)	(1,023)	(1,355)
Yelp - $$$-$$$$	32,677***	24,016***	39,666***	32,214***	28,858***
	(685.9)	(2,128)	(1,458)	(1,430)	(1,502)
Year FE	YES	YES	YES	YES	YES
Observations	3,808	918	1,054	796	364
R^2	0.482	0.356	0.567	0.329	0.510

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Notes: Observations are individual retailers from our sample able to be matched to Yelp. Independent variables are indicators for a firm's price range in Yelp, where the excluded category is Yelp '$'. Coefficients denote the average difference in firm 'quality' corresponding to different Yelp price categories. Firm 'quality' is determined by the dollar-weighted average income of customers at a given retailer.

Table 5: Customer Churn and Volatility

	(1)	(2)	(3)	(4)	(5)	(6)
VARIABLES	T. Vol.	T. Vol.	T. Vol.	I. Vol.	I. Vol.	I. Vol.
Churn	0.0158***		0.0113***	0.0116***		0.00658***
	(0.002)		(0.002)	(0.002)		(0.002)
Observations	1,037	1,037	1,037	1,038	1,038	1,038
R^2	0.276	0.257	0.341	0.185	0.226	0.262
Specification:	Univariate	Ind. FE	Add Churn	Univariate	Ind. FE	Add Churn

	(7)	(8)	(9)	(10)	(11)	(12)
VARIABLES	CAPM β	CAPM β	CAPM β	Rev. Growth	Rev. Growth	Rev. Growth
Churn	0.783***		0.368**	0.199***		0.0860**
	(0.140)		(0.172)	(0.055)		(0.038)
Observations	1,037	1,037	1,037	1,034	1,034	1,034
R^2	0.245	0.377	0.409	0.165	0.326	0.344
Specification:	Univariate	Ind. FE	Add Churn	Univariate	Ind. FE	Add Churn

Notes: The level of customer churn is calculated at a firm-year level (2011-2014), and it is the churn from last year's customer base. "T. Vol." is total volatility, the standard deviation of daily stock returns in that year. "I. Vol." is idiosyncratic volatility, the standard deviation of daily CAPM residuals in that year. "CAPM β" is the beta from a regression of a stock's daily excess returns on the excess returns of the market in a given year. "Rev. Growth" is the absolute value of the log change in year-over-year revenue. All regressions are value weighted: each observation has a weight proportional to the firm's lagged market capitalization. Standard errors are clustered at the firm level. All LHS variables Winsorized at the 1% and 99% level. The "Ind. FE" specification includes fixed effects for the industry groups: Restaurants, General Merchandise, etc. The "Add Churn" specification keeps the industry fixed effects, and adds our churn measure.

Table 6: Customer Churn and Firm Characteristics

VARIABLES	(1) Invest Rate	(2) Profit Rate	(3) Markup	(4) Q	(5) SD(Invest Rate)	(6) SD(Profit Rate)	(7) SD(Markup)
Annual Customer Churn	-0.0246***	-0.0695***	-0.0188**	-1.885***	0.0394***	0.0782***	0.0243***
	(0.00868)	(0.0139)	(0.00772)	(0.245)	(0.00464)	(0.00620)	(0.00448)
Observations	2,319	2,362	1,943	2,082	2,349	2,349	2,349
R^2	0.121	0.049	0.041	0.122	0.044	0.070	0.020
Year FE	YES	YES	YES	YES	YES	YES	YES
Industry FE	YES	YES	YES	YES	YES	YES	YES

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Notes: Investment rate measured as the ratio between capital expenditures and lagged assets. Profitability is measured as the ratio between net income and lagged assets. Markups are proxied for by gross margins. Tobin's Q is measured as the inverse of book to market ratio. Columns 5-7 measure the time series standard deviation of a given variable scaled by the average standard deviation of that firm's Tobin's Q. The level of customer churn for each firm is calculated at a firm-year level and then averaged across all years in the sample (2010-2015). Average firm churn is then applied to all available years of Compustat data back to 1970 for each firm.

Table 7: Customer Churn and Firm Investment Dynamics

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
VARIABLES	All Firms	All Firms	All Firms	All Firms	All Firms	All Firms	Retail	Retail	Retail
Q_{t-1}	0.00730***	0.00706***	-0.00429	0.00964***	0.0104***	0.00403	0.0162***	0.0105***	0.00224
	(0.000399)	(0.00146)	(0.00406)	(0.000393)	(0.00145)	(0.00398)	(0.00126)	(0.00107)	(0.00284)
Q_{t-1}*Low SG&A	0.00857***			0.00817***			0.00168		
	(0.000568)			(0.000549)			(0.00152)		
Q_{t-1}*High Churn		0.00973***			0.00674**			0.00552***	
		(0.00257)			(0.00246)			(0.00181)	
Q_{t-1}*Churn			0.0269***			0.0131*			0.0186***
			(0.00712)			(0.00696)			(0.00497)
Observations	56,676	2,038	2,038	56,661	2,037	2,037	8,691	1,690	1,690
R^2	0.402	0.415	0.415	0.444	0.489	0.489	0.478	0.525	0.526
Firm FE	YES	YES	YES	YES	YES	YES	YES	YES	YES
Year FE	NO	NO	NO	YES	YES	YES	YES	YES	YES

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Notes: Investment rate measured as the ratio between capital expenditures and lagged assets. Tobin's Q is measured as the inverse of book to market ratio. The level of customer churn for each firm is calculated at a firm-year level and then averaged across all years in the sample (2010-2015). 'Low SG&A' ('High Churn') is an indicator at a firm-level for being in the bottom (top) of the SG&A (customer churn) distribution across firms. 'Churn' is a continuous variable measuring average firm-level customer churn. Retail firms are those with the one-digit SIC code of 5.

Table 8: Customer Churn and Revenue Decline During COVID-19 Outbreak

	(1)	(2)	(3)	(4)	(5)
VARIABLES	ln(Spend)	ln(Spend)	ln(Spend)	ln(Spend)	ln(Spend)
March 2020	-0.304***	-0.0535	-0.202***		
	(0.00895)	(0.0349)	(0.0171)		
Mar 2020*Churn		-0.375***		-0.432***	
		(0.0505)		(0.0587)	
Mar 2020*Churn Quartile			-0.0416***		-0.0298***
			(0.00590)		(0.00671)
Observations	139,089	139,089	139,089	139,089	139,089
R^2	0.910	0.910	0.910	0.915	0.915
Month of Year FE	YES	YES	YES	YES	YES
Day of Month FE	YES	YES	YES	YES	YES
Day of Week FE	YES	YES	YES	YES	YES
Firm FE	YES	YES	YES	YES	YES
Industry FE	YES	YES	YES	YES	YES
Industry*Month FE	NO	NO	NO	YES	YES

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Notes: The level of customer churn for each firm is calculated at a firm-year level and then averaged across all years in the sample (2010-2015). 'March 2020' is an indicator equal to one in March of 2020. It is interacted with the continuous measure of churn and with churn as binned into four quartiles. Spending data spans January 1, 2019 to March 31, 2020. Continuous measure of churn ranges from roughly 0.33 - 0.9.

Table 9: Customer-Base Similarity and Returns

	Low	2	3	4	High	Long/Short
MKT	1.064***	1.065***	1.071***	0.959***	0.978***	-0.086
	(0.082)	(0.082)	(0.067)	(0.069)	(0.065)	(0.085)
SMB	-0.042	0.021	-0.042	-0.065	-0.198**	-0.155
	(0.150)	(0.130)	(0.139)	(0.110)	(0.092)	(0.174)
HML	-0.207	-0.32	-0.185	-0.027	-0.184	0.023
	(0.158)	(0.197)	(0.126)	(0.166)	(0.143)	(0.178)
RMW	0.438**	0.576***	0.512***	0.273	0.256*	-0.182
	(0.187)	(0.206)	(0.175)	(0.176)	(0.131)	(0.215)
CMA	0.171	0.048	-0.307	-0.077	-0.084	-0.255
	(0.189)	(0.315)	(0.213)	(0.216)	(0.196)	(0.213)
MOM	0.154	0.019	0.236**	0.071	0.081	-0.074
	(0.100)	(0.089)	(0.108)	(0.087)	(0.092)	(0.118)
Alpha	-0.003	-0.002	0.003	0.003	0.005**	0.009***
	(0.003)	(0.003)	(0.002)	(0.003)	(0.002)	(0.003)
Obs	108	108	108	108	108	108
R-sq	0.706	0.674	0.728	0.683	0.715	0.046
Sharpe Ratio	0.664	0.739	1.133	1.074	1.353	0.832
Mkt. Sharpe Ratio	0.91	0.91	0.91	0.91	0.91	0.91

Notes: 10 closest firms, 2010-2018, exclude finance/utilities, drop 2010 and 2015 from our data vw portfolio of nearest firms, returns over the past quarter.

Table 10: Customer-Base Similarity and Earnings Reports

	SUE		Earnings Returns		Forecast Accuracy	
	(1)	(2)	(3)	(4)	(5)	(6)
Overlapping SUE	0.00728*					
	(0.004)					
Your SUE		0.0112**				
		(0.005)				
Overlapping return			0.0153***			
			(0.005)			
Your return				0.0336***		
				(0.006)		
Overlapping forecast error					-0.00427	
					(0.003)	
Your forecast error						-0.0067
						(0.006)
Observations	59,660	74,178	59,660	74,178	59,580	73,983
R-Squared	0.208	0.041	0.125	0.057	0.358	0.034

Notes: 20 closest firms, 2010-2018, drop 2010 and 2015 from our data, require firms to have same fiscal period end, and release earnings in the same calendar quarter. All specifications include calendar quarter fixed effects and firm fixed effects. Standard errors clustered at the security level.

Table A.1: Asset Pricing Application (timing)

	Low	2	3	4	High	HML
MKT	0.806***	0.881***	0.939***	0.883***	0.848***	0.042
	(0.110)	(0.062)	(0.157)	(0.112)	(0.112)	(0.143)
SMB	-0.023	-0.284*	-0.492**	-0.036	-0.13	-0.107
	(0.141)	(0.154)	(0.190)	(0.172)	(0.146)	(0.198)
HML	-0.306	-0.216	-0.411	-0.024	-0.295	0.011
	(0.236)	(0.198)	(0.408)	(0.270)	(0.223)	(0.347)
RMW	0.063	0.055	-0.15	-0.25	0.259	0.195
	(0.315)	(0.237)	(0.449)	(0.235)	(0.239)	(0.373)
CMA	0.703**	0.203	0.047	0.116	0.076	-0.627
	(0.271)	(0.299)	(0.420)	(0.367)	(0.322)	(0.438)
MOM	-0.220*	0.067	0.281*	0.083	-0.014	0.206
	(0.117)	(0.087)	(0.165)	(0.115)	(0.121)	(0.143)
Alpha	-0.002	0	0.004	0.005	0.006	0.008*
	(0.004)	(0.003)	(0.004)	(0.003)	(0.004)	(0.005)
Obs	48	48	48	48	48	48
R-sq	0.673	0.728	0.621	0.687	0.636	0.162
Sharpe Ratio	0.492	1.117	1.466	1.512	1.557	1.32
Mkt. Sharpe Ratio	1.079	1.079	1.079	1.079	1.079	1.079

Notes: 10 closest firms, 2012-2016, exclude finance/utilities, drop 2010 and 2015 from our data vw portfolio of nearest firms, returns over the past quarter.

Table A.2: Asset Pricing Application (correlation)

	Low	2	3	4	High	HML
MKT	0.842***	1.042***	0.926***	0.893***	0.946***	0.104
	(0.075)	(0.056)	(0.059)	(0.086)	(0.103)	(0.150)
SMB	0.051	-0.095	-0.053	-0.185	-0.117	-0.169
	(0.121)	(0.108)	(0.101)	(0.121)	(0.148)	(0.228)
HML	-0.209	-0.096	-0.158	-0.1	-0.568***	-0.36
	(0.136)	(0.113)	(0.141)	(0.177)	(0.210)	(0.281)
RMW	0.145	0.563***	0.771***	0.268	0.213	0.069
	(0.157)	(0.163)	(0.169)	(0.189)	(0.227)	(0.278)
CMA	0.068	-0.397**	0.154	0.137	0.591**	0.523
	(0.216)	(0.189)	(0.240)	(0.225)	(0.271)	(0.375)
MOM	0.14	0.074	0.109	0.253*	0.169	0.029
	(0.086)	(0.089)	(0.088)	(0.133)	(0.116)	(0.172)
Alpha	0.004	0	0	0.002	0.003	-0.001
	(0.002)	(0.002)	(0.002)	(0.003)	(0.003)	(0.004)
Obs	108	108	108	108	108	108
R-sq	0.642	0.761	0.693	0.608	0.569	0.036
Sharpe	1.148	0.941	0.987	1.038	1.098	0.116
Mkt. Sharpe	0.91	0.91	0.91	0.91	0.91	0.91

Notes: 10 closest firms, 2010-2018, exclude finance/utilities, drop 2010 and 2015 from our data vw portfolio of nearest firms, returns over the past quarter. 'Sharpe' denotes the Sharpe Ratio.

Table A.3: Asset Pricing Application (double sort)

	Low Overlap		High Overlap		High-Low	
	Low Corr.	High Corr.	Low Corr.	High Corr.	Low Corr.	High Corr.
MKT	0.959***	0.928***	1.000***	0.833***	0.041	-0.094
	(0.067)	(0.078)	(0.064)	(0.068)	(0.093)	(0.095)
SMB	-0.035	-0.117	-0.068	-0.14	-0.033	-0.023
	(0.112)	(0.140)	(0.112)	(0.094)	(0.176)	(0.186)
HML	-0.218	-0.342*	-0.07	-0.233	0.148	0.108
	(0.135)	(0.204)	(0.131)	(0.142)	(0.208)	(0.263)
RMW	0.377**	0.397*	0.491***	0.169	0.114	-0.228
	(0.177)	(0.203)	(0.144)	(0.173)	(0.245)	(0.270)
CMA	0.038	0.283	-0.299	0.249	-0.337	-0.035
	(0.214)	(0.305)	(0.196)	(0.162)	(0.312)	(0.368)
MOM	0.135**	0.154	0.09	0.214***	-0.044	0.061
	(0.064)	(0.120)	(0.090)	(0.074)	(0.110)	(0.141)
Alpha	-0.002	-0.002	0.004*	0.005**	0.006*	0.007**
	(0.002)	(0.003)	(0.002)	(0.002)	(0.003)	(0.003)
Obs	108	108	108	108	108	108
R-sq	0.734	0.566	0.732	0.693	0.021	0.018
Sharpe	0.748	0.666	1.19	1.429	0.65	0.627
Mkt. Sharpe	0.91	0.91	0.91	0.91	0.91	0.91

Notes: 10 closest firms, 2010-2018, exclude finance/utilities, drop 2010 and 2015 from our data vw portfolio of nearest firms, returns over the past quarter. 'Sharpe' denotes the Sharpe Ratio.

Table A.4: Matching to Largest Firms by Industry

	Avg. Rank		Avg. Percentile Rank		% of Top 5	
Industry	Matched	Unmatched	Matched	Unmatched	Matched	Unmatched
Airlines	6	15	73%	32%	100%	0%
Clothing & Shoes	19	21	52%	48%	100%	0%
Consumer Telecom	20	66	84%	45%	80%	20%
Entertainment	11	24	77%	45%	40%	60%
General Merchandise	69	103	59%	39%	100%	0%
Groceries	6	10	58%	18%	100%	0%
Hotels, Rentals	16	32	73%	43%	60%	40%
Others Services & Tech	95	195	74%	47%	20%	80%
Resturants	30	82	76%	34%	100%	0%
Utilities	23	77	83%	43%	60%	40%

Notes: We rank compustat firms based on their total revenue in 2014. We then compare the numerical ranks (with one being the highest), and percentile ranks (with 100% being the highest) of the firms in our matched sample, with Compustat at large by industry. We then keep the 5 largest firms in each industry by revenue, and count how many of those firms are in our matched dataset. When matching to Compustat, and calculating the ranks, we restrict the sample to U.S. firms, with a traded common stock, non-missing revenue and non-missing NAICS industry.