Customer Churn and Intangible Capital*

Scott R. Baker† Brian Baugh‡ Marco Sammon§

October 2021

Abstract

Intangible capital is a crucial and growing piece of firms' capital structure and helps to explain long run trends in concentration and markups. We develop new firm level metrics regarding a key component of intangible capital – customer churn – using a class of household transaction data that is increasingly available to researchers around the world. We show that customer attachment is associated with higher markups and market to book ratios and can help to explain many dimensions of firm-level volatility and risk in both real outcomes and asset prices. This new measure provides a clearer picture of firms' customer and brand capital than existing metrics like SG&A, R&D, or advertising expenditures and is also available for private firms. We demonstrate that low levels of customer churn push firms away from neoclassical investment responsiveness and that low churn firms are better able insulate organization capital from the risk of key talent flight.

JEL Classification: D22, E22, G32, L11

Keywords: customer base, customers, transaction data, customer churn, intangible capital

*The authors wish to thank Lorenz Kueng, Steve Davis, Pete Klenow, Francois Gourio, Nicolas Crouzet, and seminar participants at the Hoover Institution, San Francisco Federal Reserve, the Kellogg School of Management, University of Pittsburgh, Rice University, the University of Delaware, West Virginia University, and NBER for their helpful comments and suggestions. An earlier version of this paper was circulated with the title 'Measuring Customer Churn and Interconnectedness'.

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

‡University of Nebraska bbaugh2@unl.edu.

§Harvard Business School mcsammon@gmail.com.

1 Introduction

Intangible capital has become an increasingly important factor of production and driver of growth for a range of industries.[1] Moreover, the growth in intangibles has been put forth as one major factor affecting the changes in corporate concentration and markups over the past several decades.

This trend has made the measurement of intangible capital more important when trying to understand firm investment and even firms' value (see e.g., Peters and Taylor (2017), Eisfeldt et al. (2020)). Intangible capital, however, is made up of a number of different components such as (1) technology/patents (see e.g., Kogan et al. (2017)), (2) customer base/branding (see e.g., Gourio and Rudanko (2014), Belo et al. (2019)), (3) human-resource intangibles (see e.g., Eisfeldt and Papanikolaou (2013)), and (4) organizational design.[2]

While firm book values may not provide a reliable guide (see e.g., Lev (2000)), a typical approach to calculate the market value of intangible capital is to capitalize spending on factors that are thought to generate intangibles. In particular, Eisfeldt and Papanikolaou (2013) create a stock of capitalized Selling, General and Administrative Expense (SG&A) as a way to measure organization capital while Peters and Taylor (2017) create a stock of capitalized R&D spending to measure intangible capital. An alternative to capitalizing expenditures is to use a 'residual' method, which attributes to intangibles all the value that cannot be explained by tangible assets (see e.g., Ewens et al. (2020)). SG&A is likely related to intangible value because unlike Cost of Goods Sold (COGS), it is spending that doesn't go into producing particular goods, but instead goes to things that build organization capital. As discussed in Lev et al. (2009), "SG&A expenses include outlays related to ... information systems, employee training, research and development, consultants' fees, and brand promotion."

Given how broad of a category intangible capital or observed SG&A spending is, it's not always obvious how these metrics should be related to firm risk or decision-making. In this paper, we use new data regarding customer behavior to pry open the black box of SG&A spending and find a component of intangible capital that has less ambiguous effects on firm riskiness and behavior. Specifically, we propose a new way to measure the value of a firm's customer base by directly

[1] See work such as Crouzet and Eberly (2019), Eisfeldt and Papanikolaou (2014), Eisfeldt et al. (2020), Ewens et al. (2020), Belo et al. (2019), Sim et al. (2013), Corrado et al. (2009)

[2] See e.g., Lev (2000), Lev and Radhakrishnan (2003) and Lev et al. (2009) for more on this decomposition.

observing levels of customer churn over time using household financial transaction data.

While we are not the first paper to explore the value of having a sticky customer base, we are the first to measure customer retention and attrition directly across a range of firms.[3] Other papers generally measure brand value indirectly by observing spending on SG&A and advertising, but we can directly measure whether or not firms do a good job of retaining a consistent customer base. Moreover, this method can be used for both public and private firms, at high frequency, and for firms without any intellectual property (e.g., patents) or intangible assets on their balance sheet.[4]

Overall, this paper makes four contributions. First, we demonstrate that household transaction data can create accurate customer-centric metrics that explain firm decisions, revenue fluctuations, and asset prices above and beyond existing indicators. This paper focuses primarily on one such metric but proposes and demonstrates others. 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.[5] 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.

By transforming household financial transaction data into firm-level panels, research in areas like industrial organization, marketing, asset pricing, and corporate finance may benefit from this class of data. Researchers have previously attained internal customer metrics for single firms, but customer data that span a diverse set of firms over time have not been readily available. Other customer-centric databases such as the Nielsen Consumer Panel prohibit researchers from de-anonymizing firm-level identifiers to link customer behavior with external firm indicators or outcomes. To our knowledge, Agarwal et al. (2020) and Klenow et al. (2020) are perhaps the only

[3]For instance, both Belo et al. (2019), Gourio and Rudanko (2014), and Morlacco and Zeke (2021) note that customer or brand capital can be a substantial portion of intangible value for some firms.

[4]Measures of intangible capital for private firms have become increasingly important, as Doidge et al. (2017) document the decline in new public listings in the US and an increase in delistings as public firms are taken private.

[5]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).

other papers that take a similar approach. They demonstrate that disaggregated spending data can provide high quality signals about consumer demand, firm growth, and equity prices for consumer-facing firms.

We demonstrate that the picture we obtain from our household transaction database yields an accurate picture of the customer base of a given firm, comparing financial and geographic characteristics for matched firms within our data to data obtained from existing external sources. Using our financial transaction data, we can accurately predict firm revenue levels and growth rates when compared to data from Compustat. We also compare the geographic distribution of revenue for firms that we observe in our data to an external measure of the geographic spread of firms' establishments, finding a close correlation between the two. Finally, we show that the average income of customer bases predicts firm prices.

A natural 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, and wholesale trade do not typically transact 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 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.

Our second contribution is to build a novel measure of customer-base churn within these firms using a transaction-level database that covers debit and credit card spending across approximately two million users. This new measure, while related to other measures of intangible capital such as brand valuation, is quite distinct from existing firm-level data or other commonly used measures of intangible capital such as SG&A spending. The marketing literature has long discussed the importance of customer base churn, but data suitable for systematic firm-level analysis has been lacking.[6]

We demonstrate that, consistent with theoretical evidence in Gilchrist et al. (2017) and Gourio and Rudanko (2014), churn is related to systematic risk. Firms with higher levels of customer

[6]For instance, Ascarza (2018) discusses the importance of targeting customers likely to churn and for whom interventions are the most effective. Lemmens and Gupta (2020) notes the range of approaches from firms to enhance customer retention, while Oded and Srinivasan (2008) highlights evidence from an alumni network categorizing customers by attachment and likelihood of churning out of a donation network.

churn tend to respond more strongly to macro fluctuations and crises. One reason for this is that when household income declines, people don't like to try new stores (see e.g., Baker et al. (2021)). CAPM beta is monotonically increasing from low to high churn portfolios and there is an increase in total stock volatility going from low to high churn firms. As a specific test of this finding, we show high churn firms were the hardest hit during the beginning of the COVID pandemic, even accounting for other measures of systematic risk, like CAPM beta, and seasonal patterns in spending across industries.

Our third contribution is to show that customer churn is related to firm-level valuation, markups, profits, and investment. In particular, high levels of customer attachment dampens volatility of both profits and investment. Such an effect is consistent with models 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.[7] 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.[8]

In earlier work, Gourio and Rudanko (2014) test for the presence of a relationship between firms' customer bases and these financial outcomes, relying on SG&A and advertising expenses across industries to proxy for frictions in matching customers to firms' products. They argue that high SG&A is indicative of more frictions and present evidence that these frictions lead firms to respond less to investment opportunities. We confirm these results and demonstrate that SG&A is a poor measure of customer-firm matching frictions within the retail and restaurant sector relative

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

[8] Our findings contribute to a bigger debate on how firms acquire customers, and how firms can extract value from their customer base over the business cycle. Dou et al. (2019) show that a crucial component of customer capital arises from key talent in the firm, and financial constraints may force this talent to leave in bad times. Gilchrist et al. (2017) show evidence that firms initially build up customer capital by charging low prices, but charge high prices in bad times – at the expense of future market share – to maintain cashflows. On the other hand, Kim (2018) shows that firms decrease prices in bad times to boost cashflows. Finally, Fitzgerald and Priolo (2018) show that firms acquire market share through SG&A, rather than through setting markups. This debate could partially be caused by measurement issues – these papers are forced to measure customer-bases indirectly, through markups and SG&A – while we have a direct measure of customer turnover.

to customer churn.

Our final contribution is to clarify the impacts of elements of intangible capital on firm-level risk. If increases in SG&A are embedded in employees (e.g., capitalized salaries and employee training), firm level risk may be elevated due to the potential for employees to take their human capital and exit a firm in order to start or join a competitor (see e.g., Eisfeldt and Papanikolaou (2013)). This makes firms which have a large stock of organization capital risky, because if the right opportunity arises, it can and will leave the firm in the form of employee attrition.

However, part of organization capital may be specific to the firm (e.g., capitalized advertising, brand promotion, loyalty programs). Unlike the part of organization capital that is specific to employees, it is hard for employees to abscond with a loyal customer base or brand capital if they want to start a new firm. While some firms interact with an ever-changing set of customers, other firms build a durable customer base that is less prone to defecting to a new competitor.

To test this, we perform a double sort on churn and capitalized SG&A (organization capital). We find that among low churn firms, there is no relationship between organization capital and systematic risk, as measured by CAPM beta. Among high churn firms, however, the relationship is monotonically increasing from low to high organization capital. In addition, among every tercile of organization capital, there is an increasing relationship between churn and risk. This is evidence that not all SG&A is going to employees – among low churn firms, SG&A is transformed into capital/brand value rather than employee human capital. This makes exiting the firm less desirable for employees due to their diminished ability to poach customers from the firm upon leaving. Among high churn firms, however, SG&A is clearly not effective at retaining customers, which implies that such expenditures are transformed to organizational capital. Thus, high churn firms with high SG&A expenditures expose themselves to greater risk of losing human capital through employee attrition.

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 evidence that we are observing accurate pictures of firm customer bases and their characteristics. Section 4 defines customer churn and details the relationship between churn and firm-level volatility. Section 5 discusses customer churn as a component of intangible capital and how churn covaries with firm characteristics and behavior. Section 6 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 smart-phone 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 typically given the opportunity to provide the service with user-names and passwords to financial accounts 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 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. Overall, our sample under-weights the lowest portions of the income distribution somewhat (e.g., households with under $10,000 per year), but otherwise spans a similar income range as the US national distribution.[9] We also find that users in our sample are well dispersed geographically in the United States, 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

[9] Appendix Figure A.1 displays the income distribution of remaining users to that of the U.S. Census in 2014, our last full calendar-year of aggregator data.

our empirical results.

One challenge with working with aggregator data is determining the accuracy of key variables, such as income and consumption. Our ability to correctly measure income depends on whether a user has linked the bank account that receives their direct deposit paychecks. If we observe no income in linked checking 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.

To address a similar 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 directly 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 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 validation 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 matching 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.

Because we link transaction descriptions to particular firms, we are unable to utilize transactions without an associated merchant. For instance, ATM withdrawals, physical checks, and payment apps (eg. Venmo or Paypal) will not be able to be matched to a merchant. This introduces some measurement error into our transaction-based pictures of firms. However, cash transactions are a fairly small and shrinking component of overall consumer spending and checks are most typically are utilized for large financial payments like rent and car payments rather than for retail goods and services purchases that we focus on.

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. Appendix Table A.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 with that description within a user.

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 associated with 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 cleaned 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 and for whom we have a relatively complete picture of 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.[10] In addition, to supplement our set of public firms, we search the web for lists of large private firms in these sectors. We find lists from sources such as Business Insider, Forbes, and Wikipedia that enumerate the largest firms and retailers in a range of categories.[11]

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. Since there are often 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.), this yields a many to one matching between descriptions and firms.

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

[10] 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.

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

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 dissimilar 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 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. In total, we are able to assign approximately 32% of total consumer spending in our dataset to a particular firm.[12]

For industries where we have extensive coverage, like airlines, general merchandise, and groceries, we are able to match all of the five 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

[12]To illustrate, we match our public firm data to Compustat, and rank firms based on their total 2014 revenue. Appendix Table A.2 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.

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 harder to track.

Table 1 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.

We think that this matching procedure suffices for illustrating the benefits of better understanding customer churn and similarities across firms. However, for researchers interested in more fully mapping out networks of competition or the entry and exit or private firms, it may be necessary to expand the matched sample. With additional work, it is possible for researchers to substantially increase the number of matches to smaller firms within the categories that we already focus on (e.g. smaller independent restaurants and retailers).

3 Validation of Firm Matching and Spending Data

In this section, we provide some evidence that our transaction data provides a meaningful view of firm customer bases and the sources of firms revenue.

3.1 Customer Characteristics and Revenue

Our first validation test is to directly compare the official revenue data to the spending that we observe at that firm for the subset of public firms in our sample (428 of 558 firms).

We match total aggregated consumer spending for public firms in our sample to their quarterly Compustat revenue data from 2010 to 2015. Given that our cleaned sample contains approximately 1.7 million users, out of a total U.S. population of 320 million (as of 2015), we would expect that the spending we observe would make up approximately 0.53% of revenue that these firms report if all firm revenue was obtained directly from consumers located in the United States. 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 1, we plot both levels of logged spending and changes in logged spending against levels of and changes in Compustat revenue. While the absolute levels are different owing to the fact that we observe only a fraction of individuals in the economy, we find a strong correlation between our own spending data and the revenue reported by public firms in relative terms. We do a good job of matching relative sizes of firms as well as the within-firm quarter-to-quarter 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).

3.2 Geographic Locations - Chain Store Guides

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 our sample of firms. We then construct two measure of firm geographic dispersion 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 2, 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.[13]

[13] Appendix Figure A.2 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.

3.3 Firm Quality Validation - Yelp

Lastly, we examine the types of users that patronize a retailer in our data and compare this to external indicators of retailer quality. We calculate average income of a firm's customers and compare this 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 2.

Unsurprisingly, we find that firms that have higher income customer bases in our data tend to be those selling higher priced goods, on average. 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.

4 Customer Churn and Firm Volatility

We now turn to our transaction based measure of churn within a firm's customer base. Broadly, we claim that customer base attributes attainable from transaction data can add significantly to the understanding of cross-sectional heterogeneity. In particular, for consumer-facing firms, our measure of customer base attachment offers a more fundamental window into the consumer attachment to firms over time and presents a clearer metric of an important element of intangible capital: customer or brand capital.

4.1 Measuring 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.[14]

In this calculation, we require that customers are observed in our data in both years. That is, that our measure of churn is not conflating attrition from our sample with attrition from a customer base. The sample in general has very low attrition; re-computing our churn measure without this restriction has a correlation of approximately 0.98 with the restricted measure that we utilize in this paper.[15]

Figure 3 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. Some of this variation is driven by the nature of contracts and competition within these industries. For instance, an individual likely only has a customer relationship with a single electricity provider, and this likely stays constant over time. Similarly, households tend to gravitate to a single local grocery store to a larger extent than they do for other retail stores.

[14] An alternative definition of churn may include only extensive margin customer adjustments; that is, new customers arriving and existing customers leaving the firm. We construct such a measure and make it available online. Downstream results in this paper are robust to using this alternative formulation of customer churn.

[15] In Appendix Figure A.3, we plot 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.

Similarly, some customer churn may be driven by regional concentration among retailers. That is, if a retailer has no local competition in a retail category, it may be difficult for customers to patronize competitors, even if they would so desire. In Table 3, we investigate this driver of customer churn. We test whether firms that have higher levels of local categorical spending shares tend to have lower levels of churn, conditional on a range of fixed effects.[16] We find that, in all specifications, higher levels of local categorical sales dominance tend to drive significantly lower levels of customer churn. Moreover, this local sales dominance produces large increases in R^2 (eg. from column 3 to column 4).

Churn may also be driven by factors such as firm-specific loyalty programs or contractual agreements. While we cannot examine the exact contracts being signed, we can proxy for such attributes at a categorical level. We split retailers into two categories: one composed of firms who generally have longer-term contracts (Utilities and Telecom firms) and the other composed of firms that who interact with customers through one-off purchases (Restaurants, Convenience Stores, General Merchandise, Groceries, and Entertainment). The first category has substantially higher levels of churn on average: about one standard deviation higher.

Moreover, the longer-term contractual firms tend to have a much weaker relationship between local customer churn and local sales dominance. Figure 4 plots within-city churn against within-city categorical sales shares. For 'Regular Purchase' firms, cities in which a firm tends to have fewer major competitors see much lower levels of churn. In contrast, for 'Long-term Contract' firms, the local sales shares have essentially no impact on local customer churn. That is, even with ample local competition, customers are often locked into a given firm for a number of years through contractual provisions.

4.2 Customer Churn and Firm Level Volatility

Churn in firm-level customer bases over time is a key metric with which to assess customer-facing firms. Higher levels of churn in firm customer bases can be a source of risk and volatility across firms who rely on such customers for their sales. To demonstrate this, we run the following regres-

[16] Local spending shares are defined as $\frac{Spending_{icjt}}{\sum Spending_{cjt}}$ where i indexes firms, c indexes categories of spending, j indexes cities, and t indexes years.

sion:

$$Outcome_{i,t} = \alpha + \beta Churn_{i,(t-1,t)} + \text{Ind. FE} + \epsilon_{i,t} \tag{2}$$

where $Churn_{i,(t-1,t)}$ is measured based on each year's customer base, relative to the previous year's customer base. 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 4 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 only 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.

While these results suggest a strong relationship between customer churn and firm volatility, two potential issues arise from Equation 2. First, this estimation approach may be masking a potential non-linear or non-monotonic relationship between churn and risk. Second, because observations are equally weighted, the estimates may be heavily influenced by small firms. To rule out these issues, we form value-weighted portfolios of firms based on the churn in their customer base and test more formally for excess firm-level equity price volatility.

Specifically, each month we form 5 portfolios on overall customer churn. To do this, we take the average of our churn measure at the GVKEY level between 2011 and 2015. We then apply this

average to all matched monthly stock-level observations in the CRSP-Compustat merged database between 2010 and 2019, which leads to about 275 matched firms each month. Based on the observations that are present each month, we compute 5 quintiles of churn. We compute value-weighted portfolios within these quintiles, where within each month, the weights are proportional to 1-month lagged market capitalization.

Table 5 contains the results from this approach, showing that CAPM betas monotonically increase from low churn to high churn portfolios. As a result, a portfolio that goes long high churn firms, and short low churn firms (5-1) loads positively and statistically significantly on the market factor. The last row of Table 5 shows that there is an increasing, but not monotonic, relationship between total volatility (standard deviation) and churn.

4.3 Firm-Specific Revenue Declines During COVID-19

COVID-19 presented an opportunity to perform an out of sample test of how having low customer-base attachment (high churn) can drive demand-side risk for firms. Previous work, such as Baker et al. (2021), 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 (ie. high churn) are more strongly impacted by the recent COVID-19 outbreak.

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.

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 aggregated levels of daily consumer spending across merchants.[17] 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 measure of customer churn.

Table 6 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. In Column 2, we see that firms with high levels of customer churn (estimated using the 2010-2015 data) saw much larger declines in customer traffic and spending than those with low levels of churn: a firm in the top quartile of churn saw a decline in spending about three times larger than those in the bottom quartile. In Column 3, we retain a strong negative impact of churn above and beyond controls for firm-level equity betas.

While there are substantial concerns about differential treatment across different sectors of the economy during COVID (eg. some types of retailers faced more legal restrictions than others), these correlations between revenue declines and customer churn are not driven by differences across industries. Using both industry and industry by month fixed effects in columns 4 and 5, we see that the effect persists with a similar magnitude. Even controlling for the average industry-level decline in consumer spending during March 2020, high churn firms still tended to see declines in consumer spending substantially greater than among low churn firms.

5 Customer Bases and Intangible Capital

Many papers have discussed the rise in intangible capital over the past decades and how this rise can lead economists and policymakers to mis-measure things like productivity growth, competition, and markups.[18] The overall stock of intangible capital held by a firm is often measured by means of acquisition premia (e.g., Ewens et al. (2020)) or through a perpetual inventory method which aggregates flows of SG&A or R&D spending (e.g., Eisfeldt et al. (2020)).

However, intangible capital is not an undifferentiated concept: it reflects an amalgamation of a number of components such as R&D and patent holdings, advertising or brand capital, knowledge

[17] 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.

[18] A small sample of papers include: Crouzet and Eberly (2019), Eisfeldt and Papanikolaou (2014), Eisfeldt et al. (2020), Ewens et al. (2020), Belo et al. (2019), Sim et al. (2013), Corrado et al. (2009).

capital held by workers, business practices such as software utilization or novel supply chains, customer capital, and organization capital. Independent measurement of these component pieces is important as these components may not be highly correlated with one another (or even positively correlated). While overall productivity may hinge on aggregate intangible capital, other elements of firm-level risk or decision-making may crucially depend on only a subset of these types and utilizing aggregate intangible capital thus may yield biased estimates when examining the impacts of particular intangible capital components on firm-level outcomes.

Customer attachment to firms is one such important component of firms' intangible capital (see e.g., Crouzet and Eberly (2019), Belo et al. (2019)). Customer attachment also enters many discussions of market power and competition as increases in attachment can enable firms to sustain higher markups for their goods. However, this component is often hard to capture in a systematic way, even for public firms, given data limitations in common sources of firm-level data such as Compustat. Our measure of customer churn speaks directly to this element of intangible capital: higher levels of customer attachment to a firm and a brand manifest in lower levels of churn within a customer base over time. While it is not a precise measure of customer matching frictions, search costs, or merchant specific match quality, it captures elements of these important firm level characteristics.

Table 7 highlights an association between customer churn and some indicators of intangible capital both within and across industries. For instance, Columns 1 and 2 examine the relationship between customer churn and firms' book to market ratios, finding that firms with lower levels of churn command higher market values relative to their book value. Columns 3 and 4 look directly at customer churn's relation to brand values, as assessed by a private research firm.[19] We find that brand value is highly correlated with levels of customer churn, especially within industries.

Customer attachment to firms may be in part driven by unique goods or services that are able to be offered only by that particular firm. Accordingly, columns 5 and 6 shows that the patent intensity of a firm is also linked to lower levels of customer churn.[20] Finally, in columns 7 and 8,

[19]Values calculated by Brand Finance's Brandirectory which looks at components such as emotional connection, financial performance and sustainability and then applies royalty rates to calculate a capitalized brand value. Appendix Figure A.4 displays the relationship between brand value rankings and churn across a range of industry categories.

[20]We value patents using the extended replication file for Kogan et al. (2017) and take annual patent values for each of our sample years (2011-2015), scaling by market capitalization.

we note that churn is also linked to an important element of profitability of firms: markups.[21]

While customer churn is a highly significant predictor of these intangibles metrics across all specifications, only a minority of variation is explained. That is, while churn is related to other measures of firm-level intangible capital, it has particular value as a more precise measure of a *specific* component of intangible capital: customer attachment.

As one more stark example, Figure 5 displays the correlation between firm-year intangibles proxies and our measure of firm-level annual customer base churn. Splitting our sample into retail and non-retail firms, a clear picture emerges: the relationship between customer churn and organization capital is negative for non-retail firms but highly positive for retail firms. These findings are mirrored when using advertising expenses or current SG&A spending in place of accumulated organization capital. Using advertising expenses or SG&A as a proxy for customer attachment or brand capital will lead researchers to substantially different conclusions in different industries.

A similar argument exists when working to measure parts of intangible capital other than customer attachment. For many firms and industries, using aggregate levels of intangible capital can significantly misstate the importance of individual components of intangible capital. Directly observing customer churn gives researchers an additional tool to understand a more precise channel regarding firms' intangible capital and the effects on investment decisions, ability to sustain markups, and market assessments of risk.

5.1 Churn and Customer Capital

Given the importance of customer capital, we seek to apply our measure of customer churn to a framework with predictions for firms' investment behavior. If customers have frictions when shifting between firms (i.e., accounting for customer attachment), customer bases act as state variables in capital adjustment cost models and thus can affect the rate of return on any given investment. As laid out in Christiano et al. (2005), these investment adjustment costs may take the form:

$$k_{t+1} = (1 - \delta)k_t + F(i_t, i_{t-1}) \tag{3}$$

[21] Markup data are obtained from Loualiche (Forthcoming) who calculates markups using the method in De Loecker et al. (2020).

$$F(i_t, i_{t-1}) = \left(1 - S\left(\frac{i_t}{i_{t-1}}\right)\right) i_t \tag{4}$$

where each period a share δ of capital depreciates, and firms purchase investment goods, i_t, to increase the capital stock. The function $F(i_t, i_{t-1})$ describes how current and past investment is transformed into installed capital that can be utilized in the next period. The convex function $S\left(\frac{i_t}{i_{t-1}}\right)$ penalizes deviations from the prior level of investment, with $S(1) = 0$.

These adjustment costs shift firms' responses to investment opportunities away from the frictionless adjustment benchmark. Our measures of firm-level customer churn are consistent with heterogeneity in the function $S(\cdot)$ across firms, as some customer bases are more difficult to adjust than others. They are also consistent with time-series variation in $S(\cdot)$, as e.g., it may be more costly to adjust a customer base during a recession when people are hesitant to try new firms/products.

Gourio and Rudanko (2014) pursue this general line of reasoning, building a model of product market competition that features customer attachment driven by frictions in search that prevent customers from costlessly shifting between firms. This results in sticky customer bases and generates empirical implications for firm-level characteristics and behavior. They use SG&A spending to proxy for levels of frictions that will generate more stable customer bases for some industries than others.

In such a framework, firms with a higher degree of customer stickiness can be expected to have higher levels of markups and higher market to book value (Q), consistent with our findings in Table 7. Firms with high levels of customer attachment (and low customer churn), are able to extract value from their customers over time after initial investments in customer acquisition.

Moreover, such low-churn firms are predicted to feature an investment profile that is smoother over time. Customer base adjustment frictions lead these firms to adjust more slowly to new investment opportunities yielding weaker investment responses to changes in Q. High churn firms more closely approximate the frictionless adjustment benchmark in a neoclassical model wherein increases in firm productivity drive immediate increases in firm investment.

Table 8 explicitly tests these predictions. First, in Column 1 we note that firms with low levels of customer churn (high stickiness) do tend to also have high market to book values relative to other firms in their industries. In column 2, we note that customer churn is a strong predictor of more volatile investment rates over time within a firm. We then examine whether customer churn is

associated with differences in firms' responses to Q shocks and whether these predictions hold for SG&A as well. That is, we test whether the neoclassical model, in which firm investment responds quickly to changes in productivity, is a weaker fit for firms with high levels of customer attachment (and low customer churn). Explicitly required for SG&A to perform well as a proxy for customer capital is that SG&A is highly linked to firms or industries that have high barriers/frictions in their markets.

In Column 3, we show that firms with low levels of SG&A do appear to be more like 'classical' no-adjustment-cost firms who respond more strongly to shocks to Q than firms with higher levels of SG&A spending. In Column 4, we repeat this regressions, restricting the sample solely to firms in the retail sector (SIC-1 code of 5). Here, the coefficient on SG&A switches sign and is significantly different than zero, producing an effect opposite to our prediction. We assert that this change in sign is not due to this conceptual model failing among such firms, but because SG&A is not a good predictor of customer stickiness within the retail industry. Firms in this industry with the highest levels of customer attachment tend to be those that actually spend only small amounts on SG&A, as seen in Figure 5.

Columns 5 and 6 include an interaction of lagged Q with an indicator for a firm having higher than median levels of annual customer churn alongside the low SG&A indicator. Here, the interaction terms on the high churn are highly significant and of the predicted sign when examining all firms and when restricting to retailers. High churn firms tend to respond about 50% more strongly to changes in Q than do low churn firms. Moreover, controlling for firm level churn renders the coefficients on the SG&A interaction term near-zero in magnitude and statistically insignificant. In short, our measure of customer churn consistently demonstrates the impacts of firm-level customer search frictions while SG&A likely yields substantially biased estimates of the effects of customer capital, at least in a subset of industries.

5.2 Churn and Organization Capital

Separately identifying customer attachment as a component of intangible capital can not only make inferences regarding customer capital clearer, but can also clarify the impacts of other elements of intangible capital within a firm.

For instance, in Eisfeldt and Papanikolaou (2013), higher levels of organization capital (O)

make a firm riskier both in terms of total volatility of their stock returns and their CAPM beta. This is because firm i's efficiency in using O_i, ϵ_i, is set to the level of aggregate efficiency, x_t, at the time the firm is founded, τ. Efficiency follows a random walk, and if x_t becomes high, it is attractive for employees to leave and start a new firm. This is because O is specific to employees, not the firm, so they can take the stock of O with them and use it more efficiently in the new firm. This makes x_t shocks a source of risk for firms, where exposure is proportional to the level of organization capital.[22]

Our prior is there are multiple types of organization capital that have different implications for firm riskiness. One way to formalize this thinking is to modify the baseline model of Eisfeldt and Papanikolaou (2013), splitting organization capital into two components: (1) $O^{employees}$ which employees can take with them if they start a new firm and (2) O^{brand} which is specific to the firm, and thus cannot be absconded with by the employees.

In our proposed modification, $O_i = O_i^{employees} + O_i^{brand}$, so it is possibles for firms to have high O_i, but not be very risky. In particular, higher levels of O^{brand} do not expose firms to more x_t risk.[23] We believe firms with low churn have relatively more of their organization capital in brand value, while firms with high churn have relatively more of their organization capital in their employees.

To test our hypothesis regarding different types of organization capital, each month we perform a 3 $\times$ 3 sort on churn and (Organization capital)/(Total book assets plus organization capital), hereafter OK/AT. Organization capital is measured by capitalizing SG&A in a perpetual inventory method (see e.g., Eisfeldt and Papanikolaou (2013), Eisfeldt and Papanikolaou (2014), Eisfeldt et al. (2020)).[24] To this end, we first sort firms into 3 terciles of churn. Then, within each of these 3 buckets, we form 3 sub-terciles based on OK/AT. To reduce the influence of small firms, within each month, observations are value-weighted within each of the 9 portfolios.

[22] Sun and Xiaolan (2019) also embed intangible capital in firms' employees and build a model in which firms mitigate this risk through deferred employee compensation. Taking the model to the data, they proxy for intangible capital using capitalized R&D expenses.

[23] In this way, O^{brand} acts more like physical capital in the model, K, while $O^{employees}$ is exactly like O.

[24] We obtain data on organization capital scaled by total assets from the authors' GitHub repository. Following Eisfeldt et al. (2020), we remove all observations were OK/AT is 0 because SG&A is missing/zero in Compustat, or where OK/AT is less than zero because book assets are less than zero.

Table 9 contains the results. Consistent with our prior, we see that there is a monotonic increasing relationship between OK/AT and CAPM beta among high churn firms, but the relationship is nearly flat among low churn firms. We believe that this is because if a firm has both low churn, but high organization capital, SG&A is going to the firm through creating brand value. The fact that this type of organization capital is sticky means that these firms are not riskier than low churn firms with less organization capital. The opposite is true for the firms with high organization capital and high churn: their SG&A is going to employees. Because employees can leave the firm at any time, this stock of organization capital makes these firms riskier.

6 Conclusion

With the importance of intangible capital among firms growing substantially in the past decades, it is imperative to have metrics that clearly identify the components of this capital. These measures can help to illustrate the drivers of heterogeneity across industries and firms when it comes to investment, productivity, markups, and risk. Intangible capital is generally described as an amalgamation of a number of components such as brand or customer capital, organization capital, business practices, and applied R&D and patent activity. However, given data constraints, intangible capital is often proxied for through the use of capitalized SG&A spending.

Using credit and debit card transaction data, this paper demonstrates that it is possible to construct accurate pictures of firm characteristics at a highly granular level for both public and private customer-facing firms. We use this data to develop measures of firm-specific churn in customer bases that vary over time and aims to provide a tool to disentangle important elements of intangible capital across firms.

Customer churn is important for understanding both firm financial and economic outcomes. Churn correlates highly with a range of metrics of firm-level risk and volatility and outperforms such measures in predicting revenue declines during the COVID-19 pandemic. We demonstrate that churn uniquely captures elements of customer and organization capital that are unobserved when using a proxy like SG&A spending, better explaining cross-sectional variation in markups, investment behavior, and equity returns.

In addition, this paper highlights the broader potential for further customer centric measures to

be constructed with household transaction data for use by policymakers and researchers.[25] These types of indicators are possible to construct by researchers using an increasingly accessible class of financial transaction data that has been popularized by researchers in fields like household finance and macroeconomics. We would encourage other researchers in areas that focus on firm behavior and asset prices to leverage transaction data in order to answer questions regarding consumer-facing firms.

[25] Several other firm-level measures are available for download on the authors' websites.

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.

Eva Ascarza. Retention futility: Targeting high-risk customers might be ineffective. *Journal of Marketing Research*, 2018.

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. *Journal of Financial and Quantitative Analysis*, 2021.

Brian Baugh, Itzhak Ben-David, and Hoonsuk Park. Can taxes shape an industry? evidence from the implementation of the amazon tax. *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.

Frederico Belo, Vito Gala, Juliana Salomao, and Maria Ana Vitorino. Decomposing firm value. *NBER Working Paper 26112*, 2019.

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.

Carol Corrado, Charles Hulten, and Daniel Sichel. Intangible capital and us economic growth. *Review of income and wealth*, 55(3):661–685, 2009.

Nicolas Crouzet and Janice Eberly. Understanding weak capital investment: The role of market concentration and intangibles. *Working Paper*, 2019.

Jan De Loecker, Jan Eeckhout, and Gabriel Unger. The rise of market power and the macroeconomic implications. *The Quarterly Journal of Economics*, 135(2):561–644, 2020.

Craig Doidge, G. Andrew Karolyi, and Ren M. Stulz. The u.s. listing gap. *Journal of Financial Economics*, 123(3), 2017.

Winston Dou, Yan Ji, David Reibstein, and Wei Wu. Customer capital, financial constraints, and stock returns. *Financial Constraints, and Stock Returns (February 7, 2019)*, 2019.

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

Andrea Eisfeldt and Dimitris Papanikolaou. Organization capital and the cross-section of expected returns. *Journal of Finance*, 68(4), 2013.

Andrea L Eisfeldt and Dimitris Papanikolaou. The value and ownership of intangible capital. *American Economic Review*, 104(5):189–94, 2014.

Andrea L Eisfeldt, Edward Kim, and Dimitris Papanikolaou. Intangible value. Technical report, National Bureau of Economic Research, 2020.

Michael Ewens, Ryan Peters, and Sean Wang. Measuring intangible capital with market prices. *Working Paper*, 2020.

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

Doireann Fitzgerald and Anthony Priolo. How do firms build market share? Technical report, National Bureau of Economic Research, 2018.

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

Simon Gilchrist, Raphael Schoenle, Jae Sim, and Egon Zakrajšek. Inflation dynamics during the financial crisis. *American Economic Review*, 107(3):785–823, 2017.

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

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.

Ryan Kim. The effect of the credit crunch on output price dynamics: The corporate inventory and liquidity management channel. *Available at SSRN 3163872*, 2018.

Liran Einav Peter J Klenow, Jonathan D Levin, and Raviv Murciano-Goroff. Customers and retail growth. *Working Paper*, 2020.

Leonid Kogan, Dimitris Papanikolaou, Amit Seru, and Noah Stoffman. Technological innovation, resource allocation, and growth. *The Quarterly Journal of Economics*, 132(2):665–712, 2017.

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

Aurelie Lemmens and Sunil Gupta. Managing churn to maximize profits. *Marketing Science*, 2020.

Baruch Lev. *Intangibles: Management, measurement, and reporting*. Brookings institution press, 2000.

Baruch Lev and Suresh Radhakrishnan. The measurement of firm-specific organization capital, 2003.

Baruch Lev, Suresh Radhakrishnan, and Weining Zhang. Organization capital. *Abacus*, 45(3): 275–298, 2009.

Erik Loualiche. Asset pricing with entry and imperfect competition. *Journal of Finance*, Forthcoming.

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

Monica Morlacco and David Zeke. Monetary policy, customer capital, and market power. *Journal of Monetary Economics*, 131, 2021.

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

Netzer Oded and James M Srinivasan. A hidden markov model of customer relationship dynamics. *Marketing Science*, 2008.

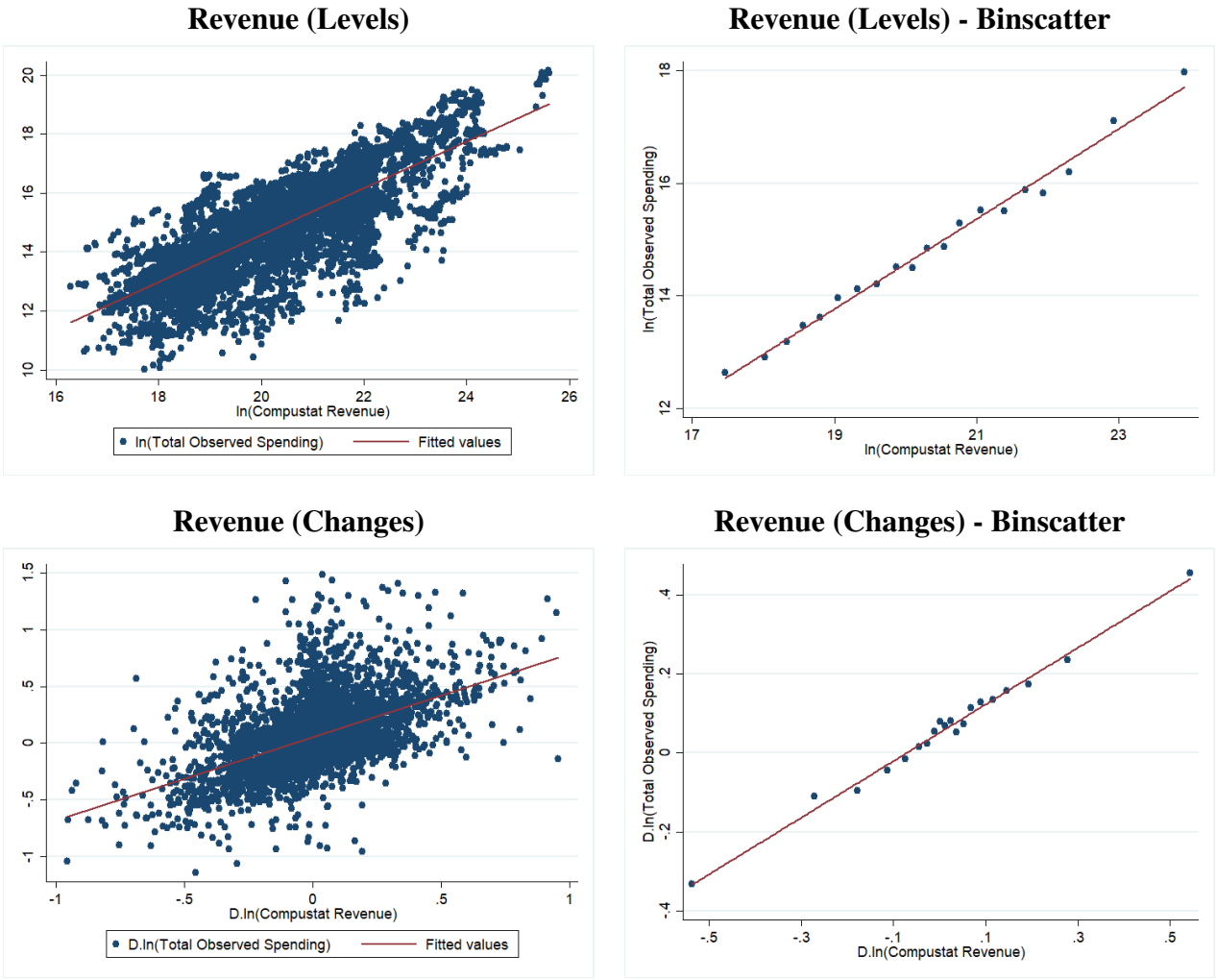
Arna Olafsson and Michaela Pagel. The liquid hand-to-mouth: Evidence from personal finance management software. *Review of Financial Studies*, 31(11):4398–4446, 2018.

R. H. Peters and L. A. Taylor. Intangible capital and the investment-q relation. *Journal of Financial Economics*, 123(2), 2017.

Jae Sim, Dalida Kadyrzhanova, and Antonio Falato. Rising intangible capital, shrinking debt capacity, and the us corporate savings glut. Technical report, Society for Economic Dynamics, 2013.

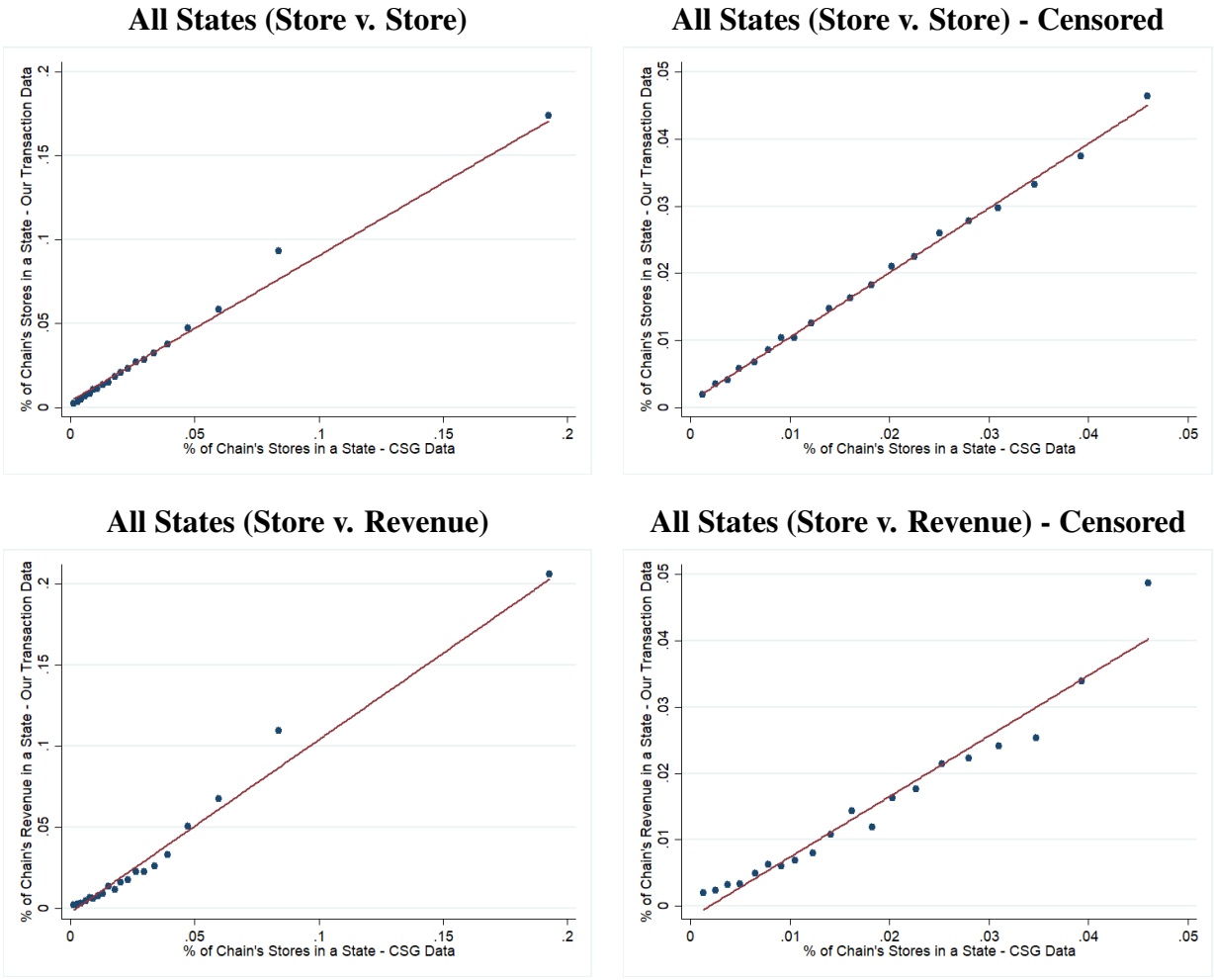
Qi Sun and Mindy Zhang Xiaolan. Financing intangible capital. *Journal of Financial Economics*, 133(3), 2019.

Figure 1: Comparison Between Reported Revenue and Observed Spending



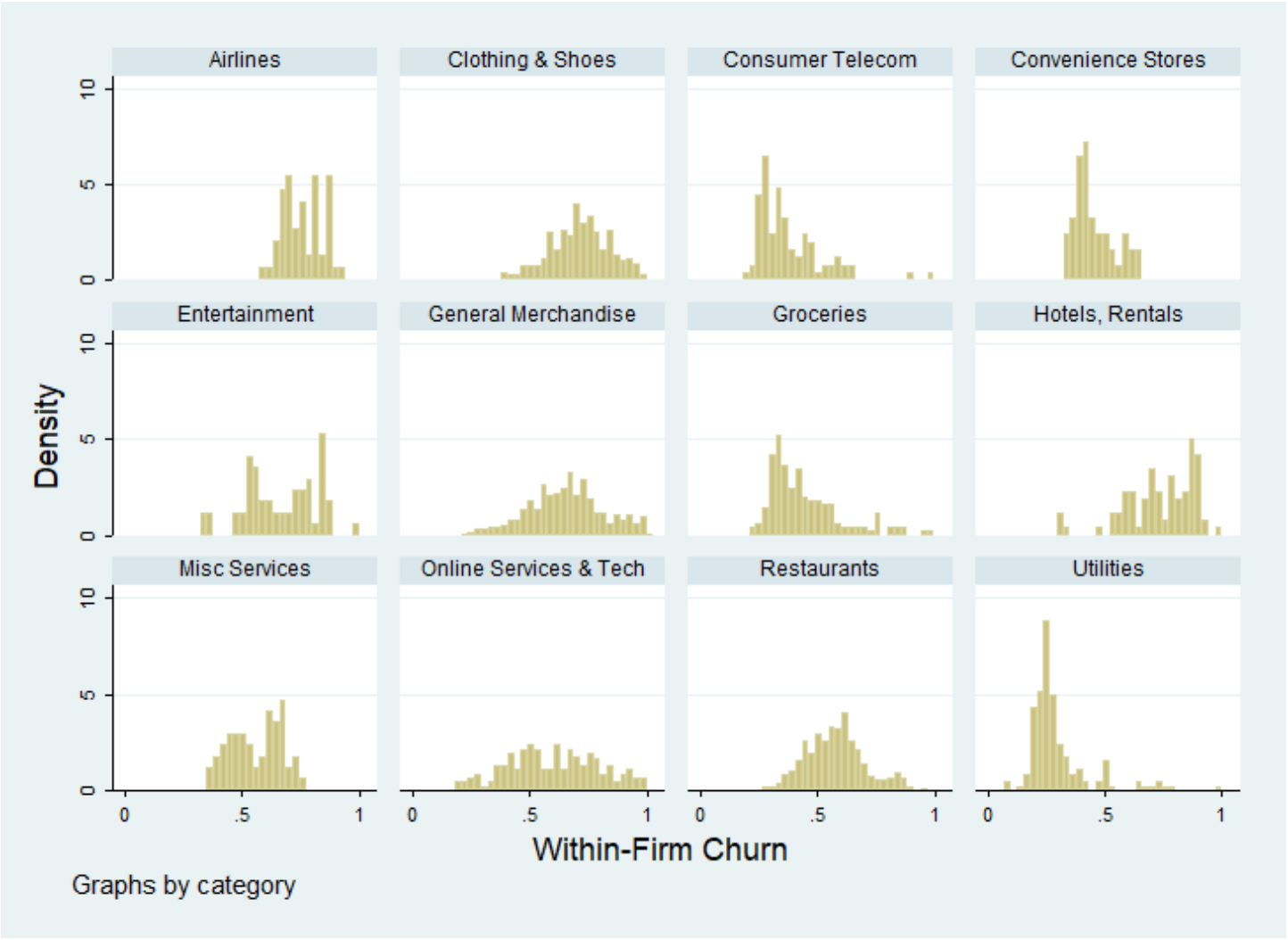
Notes: These graphs show the relationship between firm-level revenue measured in two ways: through Compustat and as observed in our transaction data. Each dot denotes a firm-quarter observation. Along the x-axis, we measure $ln(Revenue_{it})$ obtained from Compustat. Along the y-axis, we measure the 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 2: Geographic Concentration - Transaction Revenue Data and Chain Store Guide Data



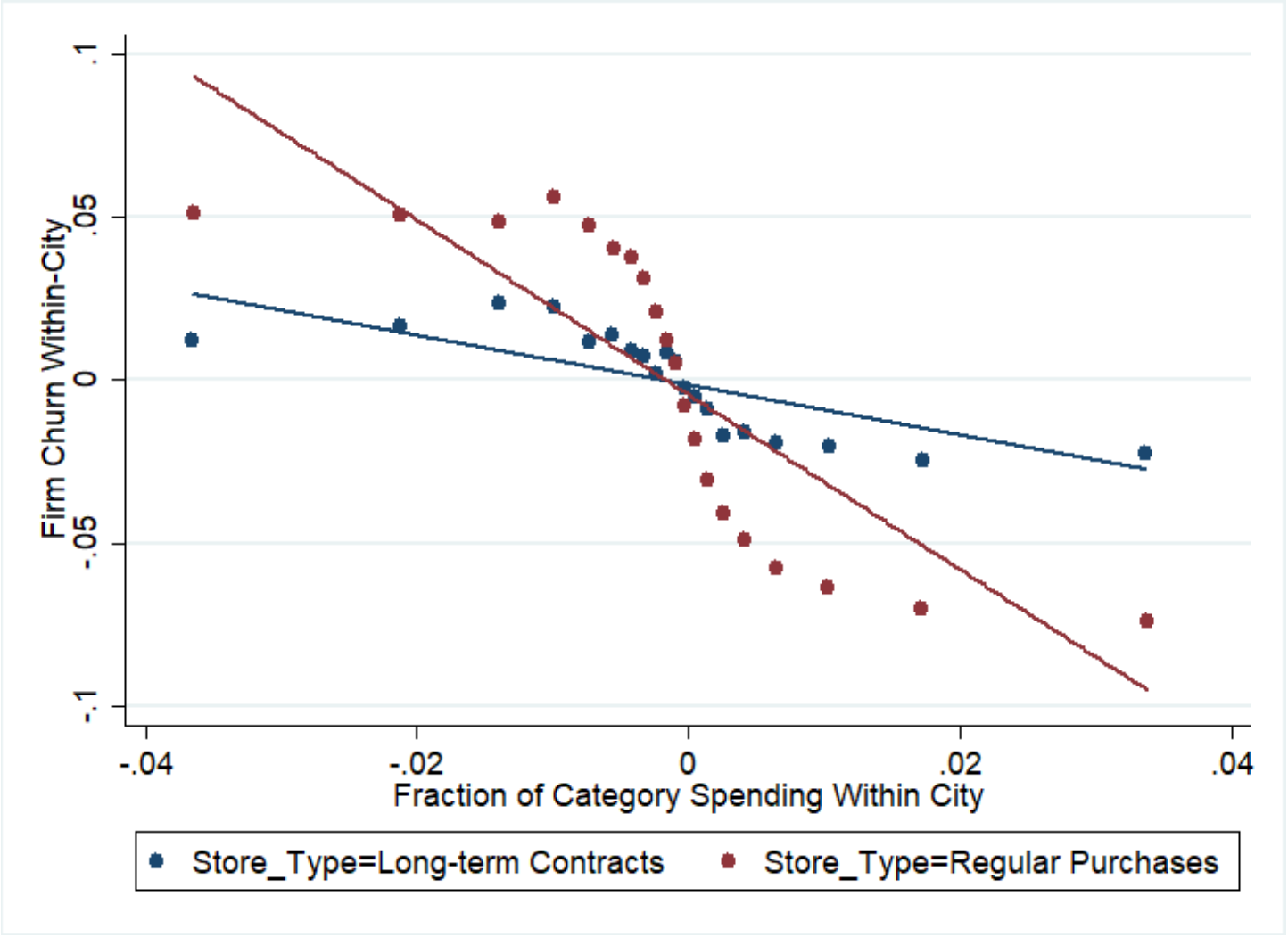
Notes: The graphs demonstrate the relationship between geographic concentration within a firm in two different ways. The first, measured on the x-axis, uses data from Chain Store Guide data and limits our sample primarily to retail firms. The x-axis measures the fraction of a firm's stores that are in 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 3: Customer-Base Annual Churn, By Industry



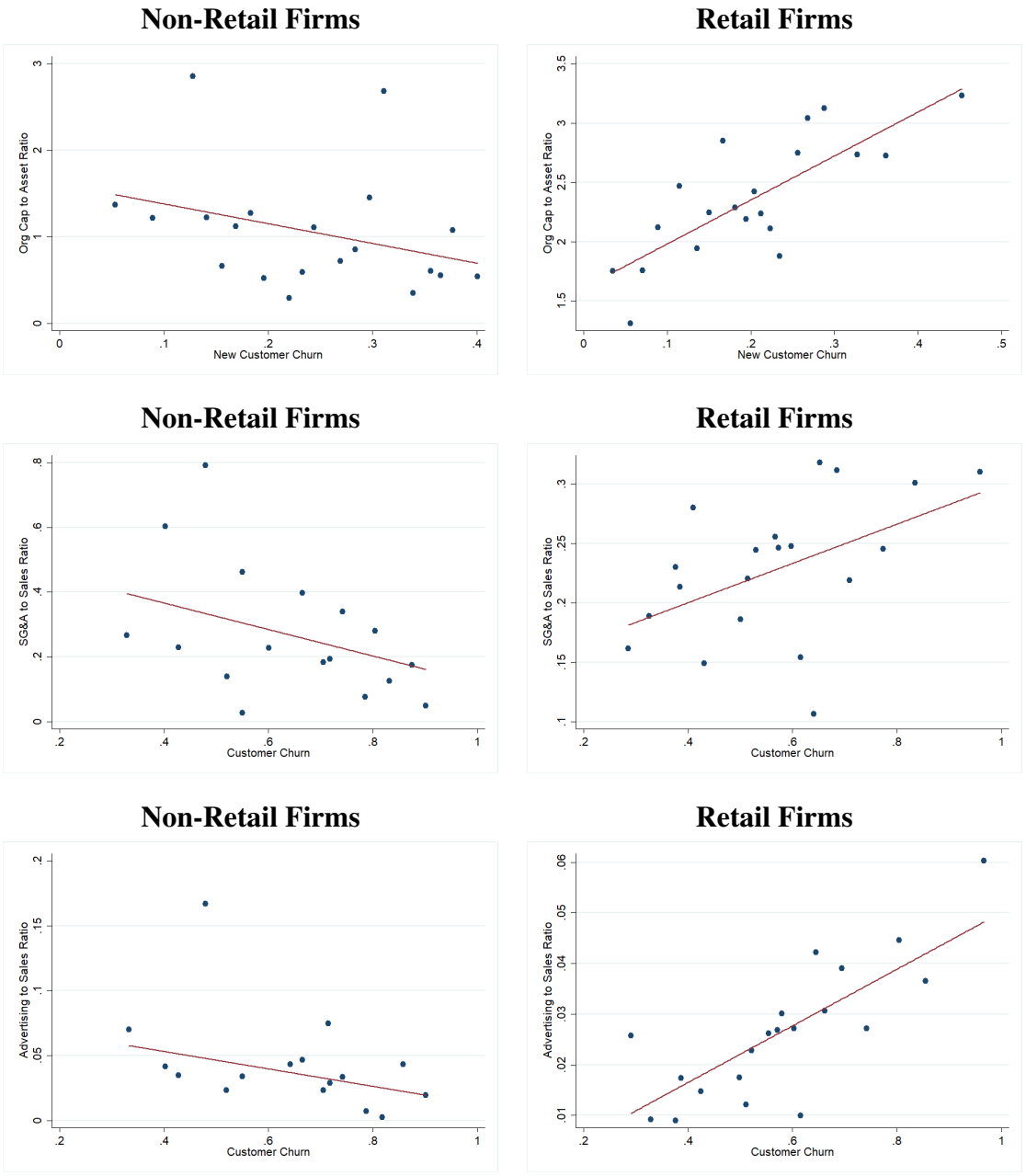
Notes: Each panel denotes the distribution of customer base churn over time across all firms in a given industry grouping in our sample. In this figure, churn is measured as the dollar-weighted overlap between the customer base of a firm f in year t and the customer base of firm f in year $t-1$. Overlap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

Figure 4: Churn and Local Sales Shares Within Category



Notes: Pictured are bin-scatter plots of churn against the fraction of spending in a category done at a given retailer. Observations are at a city-retailer-year level. Both variables are residuals of regressions on year and firm dummies. Retailers are split into two categories. The first is composed of Utilities and Telecom firms (Long-term Contracts). The second is composed of Restaurants, Convenience Stores, General Merchandise, Groceries, and Entertainment (Regular Purchases).

Figure 5: Organization Capital, S,G&A, Advertising, and Customer Churn



Notes: Retail firms defined as public firms in our sample with a one-digit SIC code of '5'. Organization Capital defined as in Eisfeldt and Papanikolaou (2013). SG&A expenses and Advertising 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.

Table 1: 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 2: 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 3: Customer Churn and Local Categorical Sales Shares

VARIABLES	(1) Churn	(2) Churn	(3) Churn	(4) Churn	(5) Churn	(6) Churn
Fraction of Category Spending in City		-0.742***		-0.566***		-0.553***
		(0.00288)		(0.00285)		(0.00285)
Observations	311,264	311,264	311,264	311,264	311,256	311,256
R^2	0.076	0.241	0.350	0.422	0.701	0.762
Year FE	YES	YES	YES	YES	YES	YES
City FE	YES	YES	YES	YES	YES	YES
Category FE	NO	NO	YES	YES	YES	YES
Firm FE	NO	NO	NO	NO	YES	YES

Standard errors in parentheses

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

Notes: The level of customer churn is calculated at a firm-city-year level (2011-2014), and it is the churn from last year's customer base. Fraction of local categorical spending is computed as $\frac{Spending_{icjt}}{\sum Spending_{cjt}}$. City-firm-years are excluded if they feature fewer than 50 customers.

Table 4: Customer Churn and Volatility

	(1)	(2)	(3)	(4)	(5)	(6)
VARIABLES	T. Vol.	T. Vol.	T. Vol.	I. Vol.	I. Vol.	I. Vol.
1 churn	0.0174***		0.0136***	0.0128***		0.00793***
	(0.00194)		(0.00274)	(0.00179)		(0.00258)
Observations	1,050	1,050	1,050	1,050	1,050	1,050
R^2	0.312	0.260	0.364	0.210	0.233	0.277
Specification	Univar	Ind FE	Add Churn	Univar	Ind FE	Add Churn

	(1)	(2)	(3)	(4)	(5)	(6)
VARIABLES	CAPM β	CAPM β	CAPM β	Rev. Growth	Rev. Growth	Rev. Growth
1 churn	0.877***		0.478**	0.228***		0.130***
	(0.146)		(0.196)	(0.0546)		(0.0448)
Observations	1,049	1,049	1,049	1,046	1,046	1,046
R^2	0.285	0.378	0.424	0.156	0.240	0.268
Specification	Univar	Ind FE	Add Churn	Univar	Ind FE	Add Churn

Robust standard errors in parentheses

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

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: within each year, 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 5: Single Sort on Customer Churn

	Low	2	3	4	High	5 - 1
Mkt. Excess Ret.	0.627***	0.958***	0.983***	1.028***	1.198***	0.571***
	(0.051)	(0.069)	(0.073)	(0.057)	(0.083)	(0.099)
Alpha	0.00526***	0.00729**	0.000721	-0.00177	0.00388	-0.00138
	(0.002)	(0.003)	(0.003)	(0.002)	(0.003)	(0.004)
Observations	120	120	120	120	120	120
R-squared	0.595	0.568	0.609	0.715	0.621	0.215
St. Dev.	0.105	0.165	0.163	0.158	0.197	0.16

Notes: Each month, we form 5 value-weighted portfolios based on average churn at the GVKEY level between 2011 and 2015. We then regress the excess returns of these portfolios on the excess return of the market factor from Ken French's data library using data from 2010 to 2019. The column "5-1" represents a long-short portfolio, which goes long high churn firms, and short low churn firms. Robust standard errors in parenthesis. The last row reports the standard deviation of each portfolio over the whole 2010-2019 sample.

Table 6: 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	-0.307***	-0.136***	-0.163***		
	(0.00885)	(0.0153)	(0.0314)		
Mar 2020*Churn		-0.917***	-1.720***	-0.798***	-1.458***
		(0.0669)	(0.124)	(0.0766)	(0.138)
Observations	141,363	141,363	42,306	141,363	42,306
R^2	0.910	0.910	0.920	0.916	0.924
Month/Day/DoW FE	YES	YES	YES	YES	YES
Firm FE	YES	YES	YES	YES	YES
Month*Beta Control	NO	NO	YES	NO	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 7: Customer Churn, Brand Value, and Markups

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
VARIABLES	B-to-M	B-to-M	ln(Brand Val)	ln(Brand Val)	Patent Intens.	Patent Intens.	Markup	Markup
Annual Customer Churn	0.426***	0.532***	-10.22***	-11.36***	-0.0418***	-0.0234***	-0.246***	-0.153***
	(0.0823)	(0.0981)	(0.504)	(0.728)	(0.00372)	(0.00418)	(0.0471)	(0.0536)
Observations	4,077	4,077	1,519	1,519	2,345	2,345	3,270	3,270
R^2	0.091	0.117	0.217	0.405	0.057	0.182	0.012	0.203
Industry FE	NO	YES	NO	YES	NO	YES	NO	YES

Standard errors in parentheses

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

Notes: The average level of customer churn for each firm is calculated at a firm-year level. Brand values calculated by Brand Finance's Brandirectory which looks at components such as emotional connection, financial performance and sustainability and then applies royalty rates to calculate a capitalized brand value. Patent intensity is calculated as the value patents (using the extended replication file for Kogan et al. (2017)) scaled by market capitalization. Markup data are obtained from Loualiche (Forthcoming) who calculates markups using the method in De Loecker et al. (2020).

Table 8: Customer Churn and Firm Investment Dynamics

	(1)	(2)	(3)	(4)	(5)	(6)
VARIABLES	Q	SD(Invest Rate)	All Firms	Retail	All Firms	Retail
Avg Annual Customer Churn	-3.838***	0.0433***				
	(0.384)	(0.00686)				
Q_{t-1}			0.00485***	0.0108***	0.00764***	0.00847***
			(0.000194)	(0.000669)	(0.00101)	(0.00116)
Q_{t-1} *Low SG&A			0.00359***	-0.00272***	0.000664	6.50e-05
			(0.000296)	(0.000884)	(0.00112)	(0.00128)
Q_{t-1} *High Churn					0.00305***	0.00416***
					(0.00110)	(0.00126)
Observations	3,082	3,611	43,837	5,622	3,220	2,371
R^2	0.143	0.188	0.641	0.662	0.616	0.622
Year FE	YES	YES	YES	YES	YES	YES
Firm FE	NO	NO	YES	YES	YES	YES
Industry FE	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. Columns 3-4 measure the time series standard deviation of a given variable scaled by the average standard deviation of that firm's Tobin's Q. 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) half of the SG&A (customer churn) distribution across firms. Retail firms are those with the one-digit SIC code of 5.

Table 9: Double Sort on Churn and Organization Capital

Churn	Low	Low	Low	2	2	2	High	High	High	HML	HML	HML
OK/AT	Low	2	High	Low	2	High	Low	2	High	Low	2	High
Mkt. Excess Ret.	0.817***	0.968***	0.733***	1.053***	0.900***	1.211***	1.088***	1.332***	1.406***	0.271**	0.364***	0.672***
	(0.066)	(0.070)	(0.078)	(0.083)	(0.083)	(0.118)	(0.080)	(0.076)	(0.114)	(0.104)	(0.105)	(0.110)
Alpha	0.00794***	0.00274	0.00537*	0.00184	-0.00142	-0.00548	0.00406	-0.00128	-0.0127***	-0.00388	-0.00402	-0.0181***
	(0.003)	(0.002)	(0.003)	(0.003)	(0.003)	(0.005)	(0.003)	(0.003)	(0.004)	(0.004)	(0.004)	(0.005)
Observations	120	120	120	120	120	120	120	120	120	120	120	120
R-squared	0.547	0.656	0.491	0.571	0.497	0.494	0.585	0.732	0.535	0.052	0.103	0.205
St. Dev.	0.143	0.155	0.136	0.181	0.166	0.224	0.184	0.202	0.249	0.154	0.147	0.193

Notes: Each month, we form 3 value-weighted portfolios based on average churn at the GVKEY level between 2011 and 2015. We then form 3 sub portfolios based on organization capital over assets from the Eisfeldt et al. (2020) replication file. We then regress the excess returns of these portfolios on the excess return of the market factor from Ken French's data library using data from 2010 to 2019. The HML columns represent a long-short portfolios, which go long high churn firms, and short low churn firms, within each OK/AT tercile. Robust standard errors in parenthesis. The last row reports the standard deviation of each portfolio over the whole 2010-2019 sample.

A Other Transaction-based Measures of Customerbase Characteristics

A.1 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. (2021), 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:

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

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 A.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 A.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[26]. 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

[26]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.

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.

A.2 Market value per customer

Although our dataset only covers about 0.8% of the US population, it is still useful for estimating the total number of customers at a given firm. To do this, we start by calculating spending per customer at the firm-year level: total spending divided by the number of unique households that shopped at the firm that year.[27] Then, to get an estimate of the number of customers, we divide total sales (SALE) in Compustat by spending per customer.

An alternative method would be to scale the number of customers at the firm-year level in our sample by our coverage of the US population. With an average coverage of 0.8%, the total number of customers at each firm should be about 1/0.008=125 times as large as the number in our sample. This gives similar estimates to the 'spending per customer' method for many large retail firms e.g., Saks and Nordstrom. It also gives similar estimates for national restaurant brands e.g., Bloomin' Brands (owner of Outback Steakhouse) and Red Lobster.

This method, however, leads to substantially different estimates for firms with a significant amount of sales outside the US e.g., Tim Hortons. While most Tim Hortons locations are in Canada, they do have several hundred US locations. This means that while their customers appear in our sample, scaling up the number of customers by a factor of 125 will likely understate the true total number of customers. If the average customer, however, is similar in the US and Canada, then our 'spending per customer' method will yield accurate estimates despite our lack of Canadian coverage.

The next step is to calculate the market value per customer: the total market capitalization at the end of the year divided by the estimated number of customers in that year. Common-sense intuition suggests that market value per customer should be higher for low-churn firms. From a present value perspective, a customer should be more valuable to a firm if they are likely to continue spending there for a long period of time. Figure A.6 plots average market value per customer vs. average churn. There is a statistically significant and economically large negative

[27] From both the numerator and the denominator we exclude household-firm-quarter observations with less than $1 of total spending.

relationship between market value per customer and churn. The standard deviation of churn is ≈0.2, so a 1 SD increase in churn would increase market value by about $180 per customer.

This result is driven mostly by differences across industries: Some of the firms with the highest market value per customer are utility companies like Dominion Energy and Duke Energy as well as Telecom companies like AT&T and Verizon. Some of the firms with the lowest market value per customer firms are struggling brick-and-mortar retailers like Barnes & Noble and Sears. While the relationship is still negative when including industry fixed-effects, the magnitude of the slope is only about 1/8th as large.

B Case Study of Shifts in Churn

While firms exhibit large differences in their average levels of customer churn when compared to each other, we also note that substantial changes in firm-level churn can take place over time within a firm. JC Penny provides one case study of how measured customer churn can be affected by corporate decision-making and customer-facing policies. In 2011, Ron Johnson was appointed as CEO of the large clothing retailer, JC Penney. JC Penney had suffered from declines in sales growth in previous years and sought a change in leadership to arrest the decline.

Johnson spearheaded a drastic change in pricing at the retailer in Q1 2012, doing away with most of the deal and coupon-based pricing and instituting more consistent low prices across the store, mirroring the approach at Johnson's former employer, Apple, which eschewed coupons and deals. JC Penney's customer base reacted strongly and negatively to this change, increasing turnover substantially in the ensuing years.

In Figure A.7, we show the rate of quarter-on-quarter customer churn for JC Penney during our sample window normalized by average churn for that quarter within the 1-digit SIC industry. The red vertical line denotes the timing of the change in pricing policy. We see a large and persistent increase in customerbase churn following this change of approximately 1.5 standard deviations.

C Customer Base Overlap and Stock Predictability

C.1 Customer Base Similarity

Another aspect of firms' customer bases that we can capture with our data 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:

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

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 A.8 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.

We note that, on average, within-industry customer base similarity is higher than that across industries. That is, many users tend to disproportionately weight their spending towards a particular industry, not simply a particular firm within an industry. However, 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.[28]

[28]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,

C.2 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 Cohen 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.

Autozone, Sally Beauty, Gamestop, Office Depot, Big Lots, and Dicks Sporting Goods.

We display the results in Table A.4. 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[29]. 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.5 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.[30]

We perform several tests of the robustness and utility of this customer base overlap measure.

[29] In unreported results, we find that this is mostly coming from across-industry customer-base similarity, rather than within-industry customer-base similarity. One explanation for this may be that within-industry links are more visible to investors who do not have access to data like ours.

[30] 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.

For instance, another potential proxy for customer base overlap is correlation of stock returns. To test where we could obtain similar results simply utilizing these correlations, we compute the correlation of each pair of firms' daily stock returns from 2011-2014. In each month, we identify the 10 most correlated firms. We repeat the procedure for forming 5 portfolios as described 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. We display these results in Table A.6. There is no pattern in the alphas from low to high, suggesting that our measure of customer base overlap contains important independent information.

Despite the results in Table A.6, it's possible that our results are still related to past correlation in stock returns. To further rule out this channel, we perform a double sort in Table A.7. 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. Both hedge portfolios have statistically significant alphas, again suggesting that our results are not driven only by correlation in stock returns among firms with high customer base overlap.

C.3 Earnings Announcements

To understand the mechanism behind the results in Table A.4, we examine days where we know fundamental information about firms is released: earnings announcements.

For simplicity, 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 because this 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. 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.[31]

In Table A.8, 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 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.

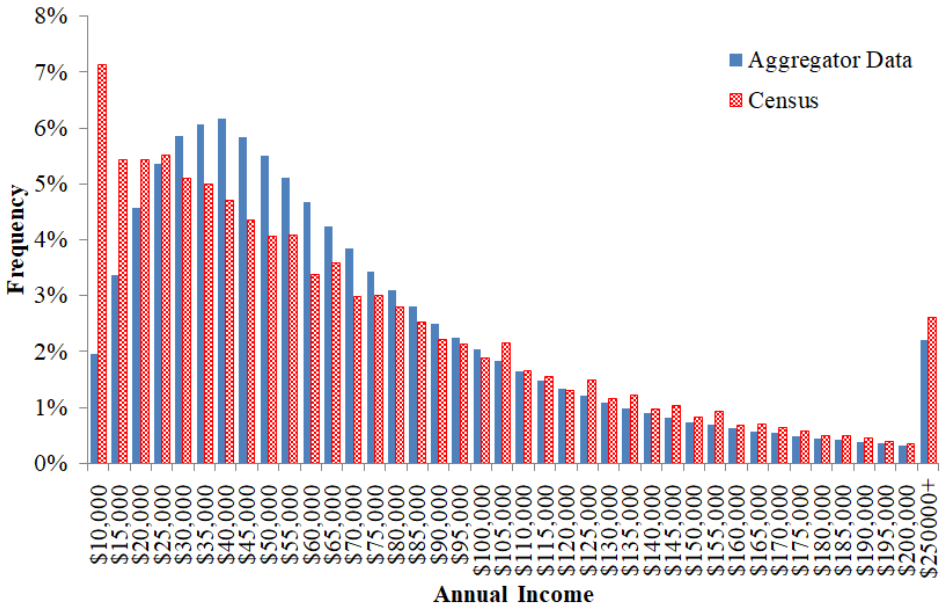
Column 3 is a regression of WMT's earnings day returns the earnings day returns of the 20 firms with the highest overlap to WMT, which released earnings before WMT in a given calendar quarter. Column 4 is a regression of the earnings day returns of the 20 firms with the highest overlap to WMT on WMT's 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.

[31] 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).

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.

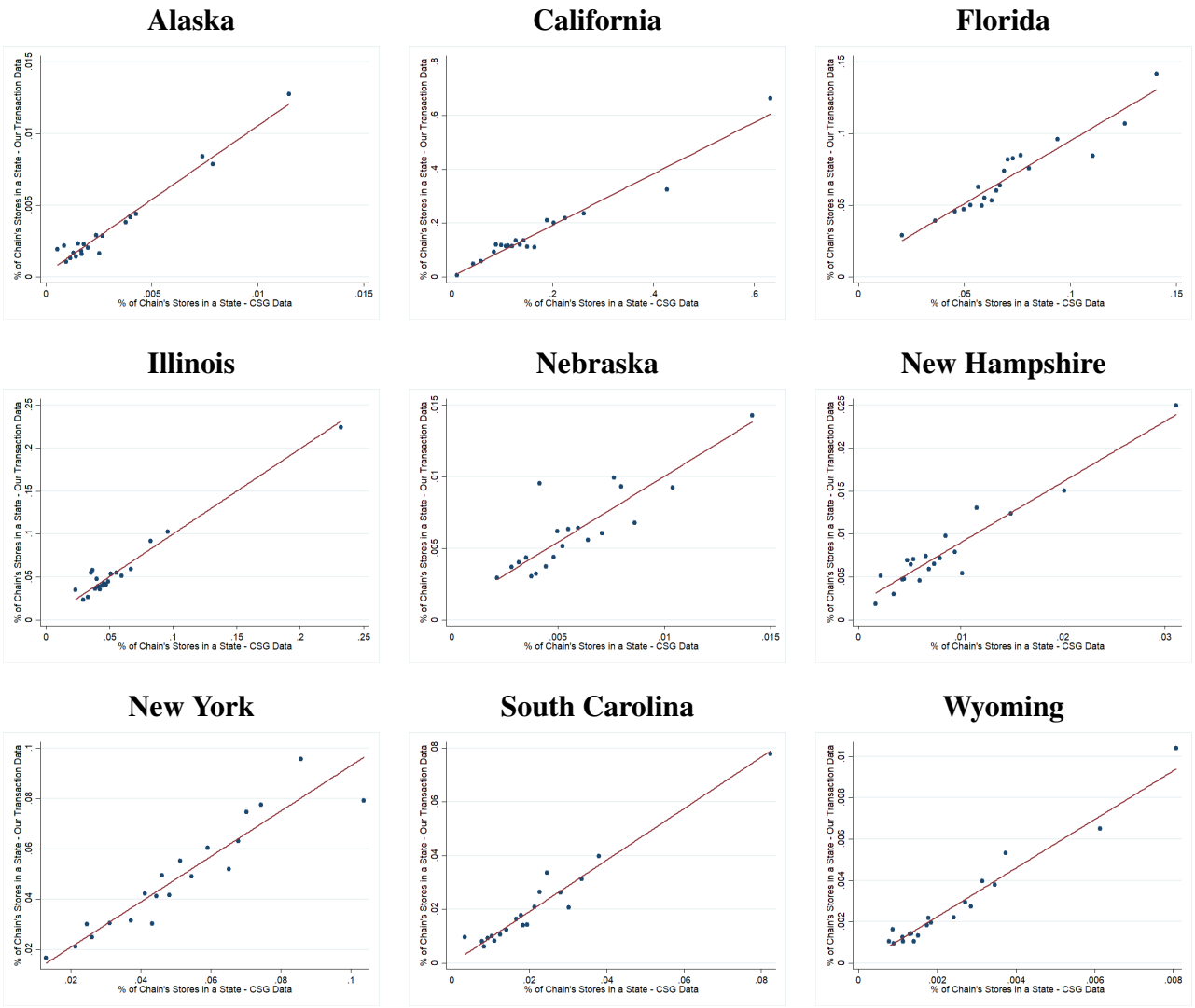
Column 5 is a regression of WMT's analyst accuracy on the analyst accuracy of the 20 firms with the highest overlap to WMT, which released earnings before WMT in a given calendar quarter. Column 6 is a regression of the analyst accuracy of the 20 firms with the highest overlap to WMT on WMT's analyst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

Figure A.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 A.2: Geographic Concentration - Transaction Store Data and Chain Store Guide Data, Selected States

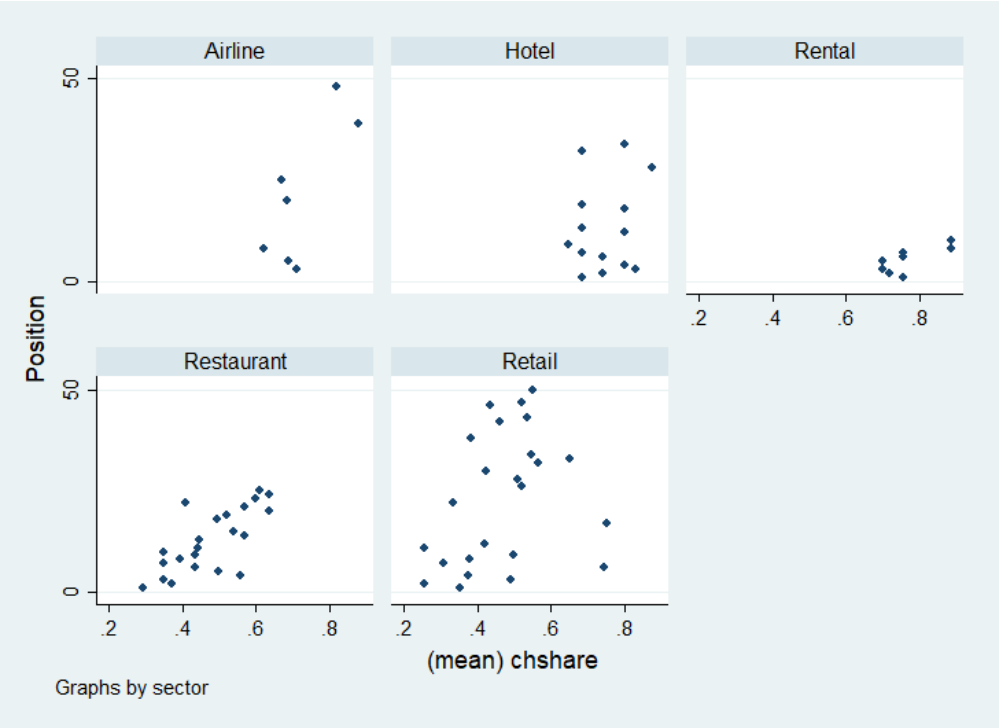


Notes: The graphs demonstrate the relationship between geographic concentration within a firm in two different ways. The first, measured on the x-axis, uses data from Chain Store Guide data and limits our sample primarily to retail firms. The x-axis measures the fraction of a firm's stores that are in 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 A.3: Customer-Base Similarity Within Firm Over Time

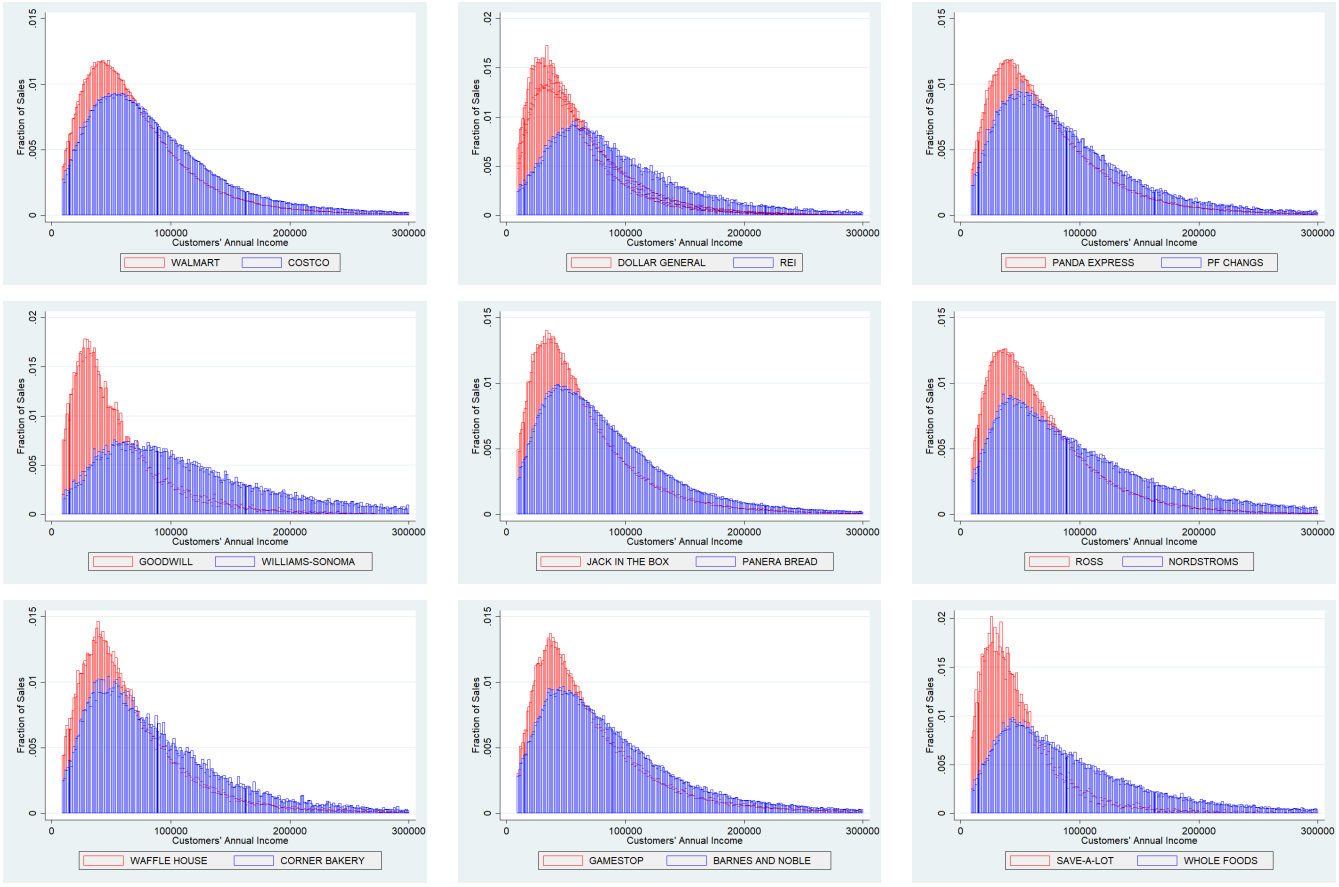
Notes: Each panel denotes the distribution of customer base churn over time across all firms in our sample. Churn is measured as the dollar-weighted overlap between the customer base of a firm f in year t and the customer base of firm f in year $t - x$ where x is between 1 and 4 and is labeled above 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 A.4: Brand Value and Churn, by Industry



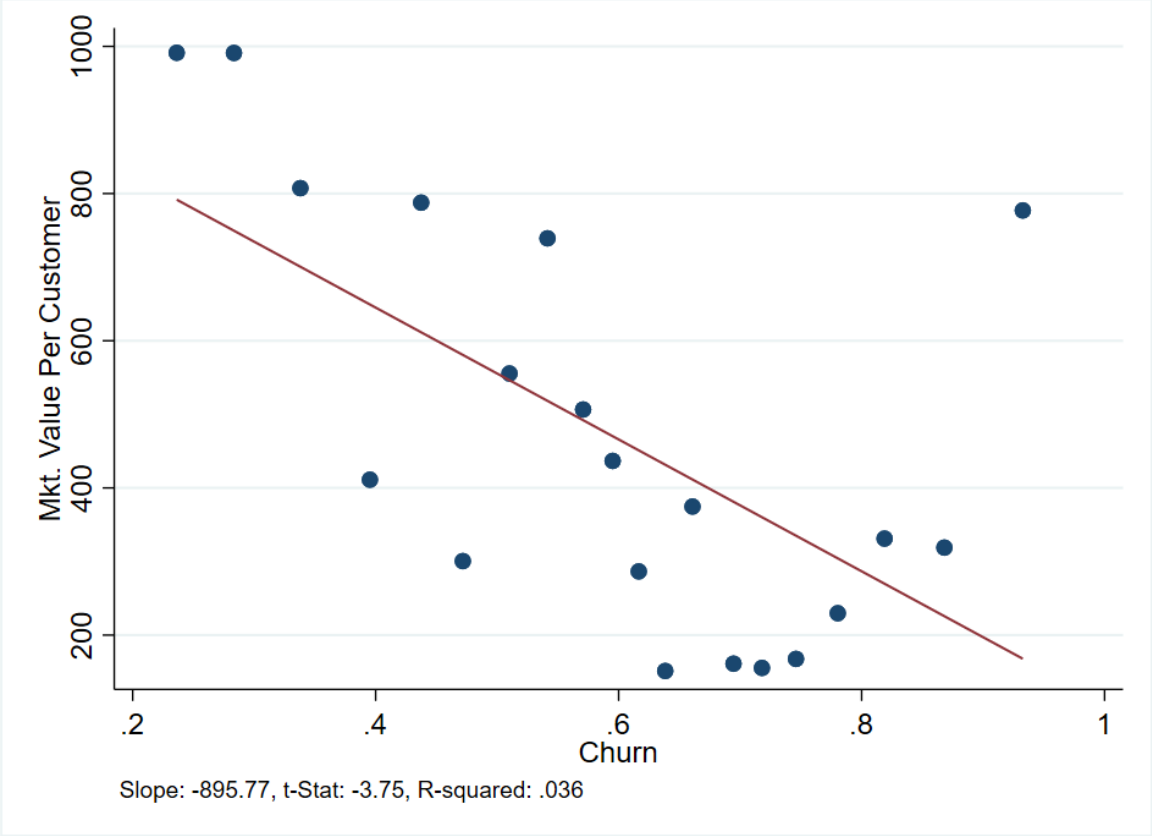
Notes: Churn denotes average annual customer churn within a firm across our sample period. Brand value rankings calculated by Brand Finance's Brandirectory which looks at components such as emotional connection, financial performance and sustainability and then applies royalty rates to calculate a capitalized brand value.

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



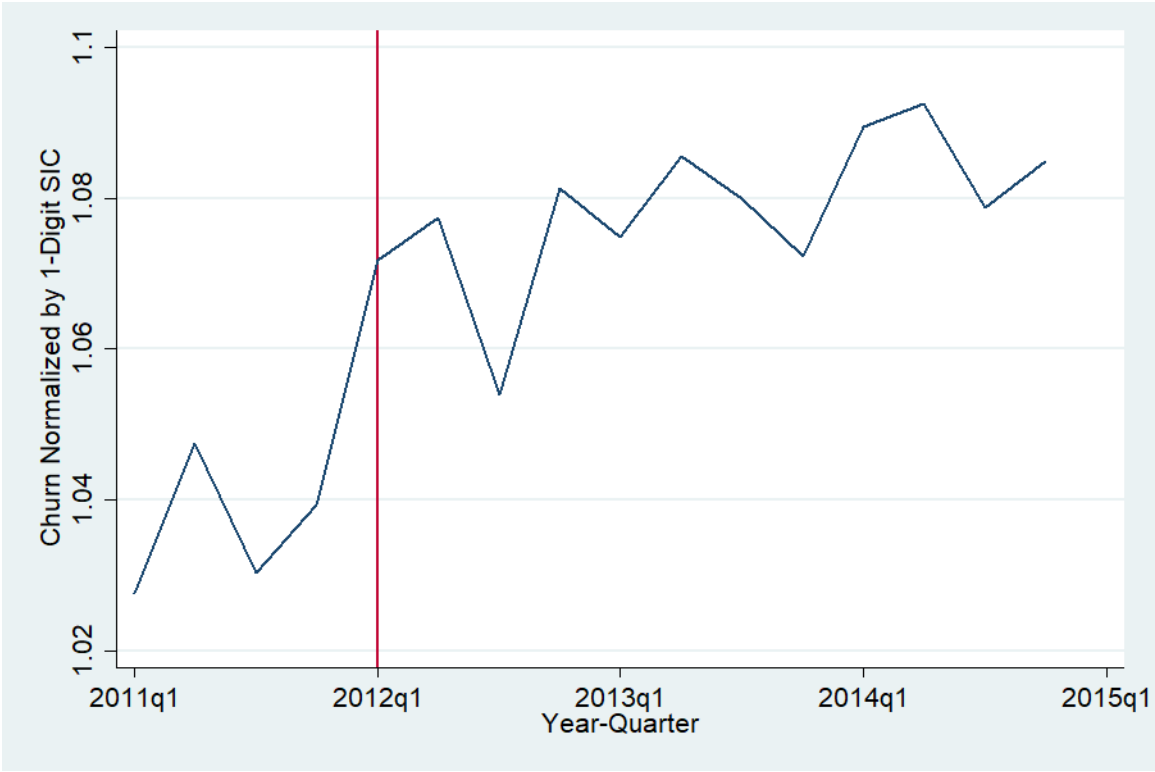
Notes: Figures demonstrate the distribution of income among customers for a selected sample of firms. Customer's are dollar-weighted by sales at a firm, so a user spending $500 at a firm will have double the weight in the histogram as a user spending $250. Annual income is binned in $1,000 increments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

Figure A.6: Market Value per Customer vs. Churn



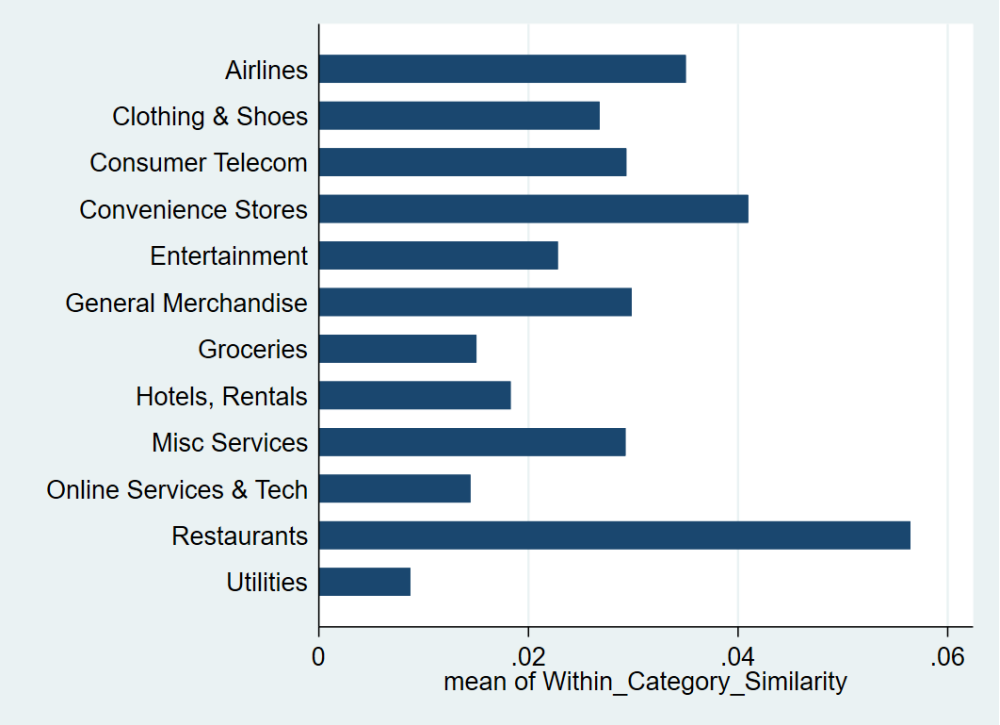
Notes: Y-axis is average market value per customer between 2011 and 2015. X-axis is average churn between 2012 and 2015 i.e., using data from 2011-2015. Estimates of market value per customer are Winsorized at the 1% and 99% level.

Figure A.7: Customer Churn at JC Penny



Notes: Plotted is the level of quarter over quarter customer churn at JC Penney normalized by the average level of quarter over quarter churn within the industry (one digit SIC code). A red line denotes the quarter (Q1 2012) in which JC Penney instituted a radical new pricing strategy.

Figure A.8: 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.

Table A.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 ave new	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
graceys 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 A.2: 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.

Table A.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 A.4: 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 A.5: 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.6: 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. Value weighted portfolio of nearest firms, returns over the past quarter. 'Sharpe' denotes the Sharpe Ratio.

Table A.7: 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. Value weighted portfolio of nearest firms, returns over the past quarter. 'Sharpe' denotes the Sharpe Ratio.

Table A.8: 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.