Information Acquisition and the Return to Data

Evidence from Firms on an E-commerce Platform

Yizhou Jin[*]
UC Berkeley

Zhengyun Sun[†]
Harvard

June 20, 2019

[New Draft with Experimental Evidence Coming Soon]

Abstract

This paper investigates online retailers' decision to acquire information and the impact of data access on their business strategy and on revenue growth. We take advantage of proprietary data from a large e-commerce platform that sells data analytics products to virtual stores operating on it. The product provides detailed information on customer sources and characteristics, aggregate demand, and competitor strategies. Our empirical investigation relies on several high-frequency panel datasets and makes use of back-end changes in the pricing, variety, and bundling of the data analytics products. Focusing on several consumer electronics and peripherals markets, we find three main results. (i) Data acquisition facilitates growth, but small retailers are very sensitive to the cost of data. (ii) Retailers take marketing and product actions with the data collected but leave prices largely unchanged. (iii) A counterfactual simulation shows that a uniform reduction in the cost of data raises overall platform sales while reducing market concentration on the margin.

[*] jyz@berkeley.edu; yjin.io.[†] zsun@g.harvard.edu.

We thank our data contacts and providers for making this project possible. They include an unnamed e-commerce platform, Carolina Harvey, and Susan Athey. We also thank Harvard Business School and Harvard Economics Department for generous financial support. For especially valuable comments and discussion, we thank Jie Bai, Ron Berman, Robin Lee, Ariel Pakes, Yuan Tian, Dennis Yao, and various seminar participants. Ability to publish is not contingent on results. Please email carolina_harvey@harvard.edu for questions about the data usage agreements.

Introduction

There is a growing consensus among business leaders that data collection and data-driven decision-making have become prevalent among firms and crucial to growth. Small-scale surveys among executives of large corporations point to rapid growth in the adoption of big data projects. 73% of respondents in a 2018 survey focusing on financial services firms report measurable realized gain from big data initiatives.[1] Researchers have also documented rapid adoption of data-driven practices and positive effect on performance and on productivity among large corporations (Brynjolfsson, Hitt, and Kim 2011; Saunders and Tambe 2015; Brynjolfsson and McElheran 2016; Wu, Hitt, and Lou 2017). Efficient methods for large firms to collect and use data have also been proposed (Dubé and Misra 2017; Nair, Misra, Hornbuckle IV, Mishra, and Acharya 2017; Bajari, Chernozhukov, Hortaçsu, and Suzuki 2019). The growing importance and sophistication of data analytics are particularly pronounced in the retail sector, largely driven by the explosive growth of e-commerce platforms. In 2018, 23% of consumer retail in China and 9% in the US happened online.[2] Consumers' and firms' behaviors online can often be tracked and quantified much more cheaply, precisely, and comprehensively.

However, much about the role of data analytics in retail businesses remains unknown. It is unclear what types of data the average retailer collects, to assist what kinds of decisions and strategies, and ultimately, to what effect. These questions become especially puzzling among small and young firms that are rarely captured by small-scale surveys used in existing literature. Social planners and platforms, on the other hand, often have incentives

[1] http://newvantage.com/wp-content/uploads/2018/02/Big-Data-Executive-Survey-2018-Findings.pdf. The survey primarily covers corporations in the financial services and insurance industry (77.2%). Another 2017 survey covering a broader range of industries and firm sizes reports that 53% of respondents say that their firms "use big data today," up from just 17% in 2015. (https://www.microstrategy.com/getmedia/cd052225-be60-49fd-ab1c-4984ebc3cde9/Dresner-Report-Big_Data_Analytic_Market_Study-WisdomofCrowdsSeries-2017.pdf).

[2] According to Euromonitor International

to facilitate growth and improve selection among exactly these firms. Existing evidence shows that, retailers of all sizes seem to be slow in responding to changes in market conditions or information technology and instead follow simple pricing strategies (Ellickson and Misra 2008; Conley and Udry 2010; Illanes and Moshary 2015; Hitsch, Hortacsu, and Lin 2017; Arcidiacono, Ellickson, Mela, and Singleton 2016; Atkin, Chaudhry, Chaudry, Khandelwal, and Verhoogen 2017). For small retailers, this is exacerbated by financial exclusion (Tybout 2000; Hsieh and Klenow 2009) and lower managerial skills in general (McKenzie and Woodruff 2015; Bloom, Mahajan, McKenzie, and Roberts 2010). Given these findings, can social planners and platforms facilitate growth among firms by shaping their data access? If so, what types of information should be provided to what types of firms, and at what costs?

In this paper, we empirically investigate the return to data access for online firms and their data acquisition decision, and in doing so, shed light on how e-commerce platforms should design and price data. Our empirical setting is a large e-commerce platform that hosts online stores. The platform offers a sophisticated data analytics product for stores operating on it. It includes detailed information on customer sources and demographics, aggregate demand in terms of transactions and search, as well as competitor strategies. Most of the information provided is proprietary to the platform. Firms can gain access by purchasing subscriptions.

The paper begins by introducing the platform and the firms operating on it. We then describe the data analytics product offered by the platform: the kinds of proprietary data that are offered, and the way they are presented and sold to firms. Next, we narrow our focus down to a set of electronics and appliances industries that have high level of e-commerce penetration.[3] We documents four key reduced-form facts. (1) Growing adoption of data

[3] 47% of all electronics and appliances are purchased online in China. See https://www.pwccn.com/en/retail-and-consumer/publications/global-consumer-insights-survey-2018-

analytics among stores, especially for information on own customers and aggregate demand; (2) small stores are very sensitive to the cost of information. Costly information acquisition is concentrated among large incumbents; (3) firms acquire costly information when they implement product and marketing strategies. We do not find evidence of that prices (in the product market) are influenced by the data collected, or vice versa; (4) information acquisition makes product and marketing strategies more effective, bringing significant growth, especially among small firms.

In uncovering these key facts, we utilize detailed data on the adoption and usage of different pieces of information contained in various modules of the data analytics tool. This is matched to panel data on the stores' operating statistics and records of various strategies implemented. Identification is achieved with back-end changes in the pricing of different data modules, or the cost of acquiring data.

In the third and fourth sections, we propose a bundling model to synthesize the reduced-form facts and to further unpack the complex correlations between stores' decision to implement a low cost marketing strategy, changing product titles, and that to acquire different pieces of information included in the data analytics products. To identify key parameters, we take advantage of the panel nature of the data, corresponding back-end changes in how data are bundled into products, as well as similar price changes as last section. We find large variation in stores' sensitivity to the cost of information. Small stores are much more sensitive, but conditional on size, young stores are less sensitive overall. Across various types of information, aggregate demand information primarily facilitate growth only among large incumbents, while competitor information seem particularly useful for young firms.

With the estimates from the structural model, we are able to conduct counterfactual

china-report.

simulations. Without taking a stand on the objective function of the platform, we only trace out the marginal impact of changing the cost of data on overall sales across all stores on the platform and on concentration. We find that a uniform reduction in the cost of acquiring data increases overall sales on the platform while reducing industry concentration on the margin.

Related Literature Our investigation contributes to several literature. First, we are related to studies on the value of information and information technology, particularly among small businesses and in developing country marketplaces. There is a large literature focusing on the impact of information and communications technology (ICT) infrastructure improvement, especially high speed Internet: they promote revenue and productivity growth and improve labor market outcomes (Cole and Fernando 2012; Hjort and Poulsen 2019). Studies on the role of information often focus on exogenous shocks to the availability of public information that is accessible to all businesses and consumers (Jensen 2007; Bai 2018). Our study contributes to the literature by modeling firms' data acquisition decision. We also quantify the growth impact of different types of information through marketing and product strategies.

We also contribute to an emerging literature on endogenous information acquisition and information design in industrial organization and quantitative marketing (Jin 2005; Dubé, Fang, Fong, and Luo 2017; Bergemann and Morris 2019; Jin and Vasserman 2019). The crucial difference is that we do not model demand or firms' competitive environment, nor do we make any structural assumptions about their priors along specific dimensions of knowledge that may be profit-relevant. Instead, we look at *marginal* changes to firm strategy, growth, and market concentration induced by endogenous information collection. Some of our findings, such as the effectiveness of product title adjustments as a marketing strategy, also echoes empirical evidence from in the digital marketing literature about the importance of search frictions, consumer salience, and obfuscation (Ellison and Fisher Ellison 2005; Ellison and Ellison 2009; Blake, Moshary, Sweeney, and Tadelis 2018).

Third, we are related to an emerging literature on behavioral economics among firms, in which firms depart from the "profit-maximizing paradigm" (Armstrong and Huck 2010). Our seemingly surprising finding of price stability, in fact, adds to a growing amount of empirical evidence on "rule-of-thumb" or cost-based pricing among offline retailers (Eichenbaum, Jaimovich, and Rebelo 2011; McShane, Chen, Anderson, and Simester 2016; Arcidiacono, Ellickson, Mela, and Singleton 2016; Illanes 2017; DellaVigna and Gentzkow 2017).[4] Meanwhile, viewing the data analytics product as a form of new technology, we also expand the literature on the resistance to beneficial new technology, particularly among small firms (Foster and Rosenzweig 1995; Conley and Udry 2010; Hanna, Mullainathan, and Schwartzstein 2014; Atkin, Chaudhry, Chaudry, Khandelwal, and Verhoogen 2017). In particular, our research setting allows us to directly observe and track hundreds of thousands of online firms with high frequency, providing a rare glimpse into the heterogeneity that exists in technology adoption across different types of firms.

1 Background and Data

1.1 E-commerce Platform

Our investigation in this chapter relies on data from a large e-commerce platform in China. There are two types of store operating on the platform, B stores are "Business-to-Consumer" businesses that are operated by registered firms, often large and already established offline. C stores are "Consumer-to-Consumer" businesses that are operated by individual

[4] There is also a large literature on rational explanations to the lack of time and geographic price variations (Berto Villas-Boas 2007; Libgober and Mu 2018; Li, Gordon, and Netzer 2018).

entrepreneurs, which makes up the vast majority of stores online. Table 1 provides key summary statistics for these stores.

From buyers' perspective, B and C stores can be found from the same main platform, although B stores have their own designated platform. For our analyses, we focus on C stores due to its prevalence. In doing so, we can largely mitigate a missing data problem due to a lack of ability to observe large stores' operations offline or on other competing platforms.

Table 1: Summary Statistics of stores on the E-commerce Platform

	Non-Sole Proprietor	Sole Proprietor
Share of Stores	4%	96%
Revenue Share	45%	55%
Avg. Age (yr)	3.4	4.1
Avg. Rating ([/15])	10.6	6.3
Annual Sales (RMB)		
Mean	5.9m	336k
Median	1.4m	24k
90th Pct.	14.1m	576k
Survival Dynamic*		
After 3 Months	96%	54%
After 12 Months	90%	24%

Note: summary stats calculated based on a sample of active stores of a randomly selected day in 2018. Survival probability calculated with all newly registered stores in 2016 and 2017 and conditioned on having stayed in the market for at least one month. B stores are formally registered, own brand names, have entry requirements, need to pay service fees but get preferential treatment in terms of search rankings and promotion events.

For every store (we use store and store interchangeably), we observe a rich set of observable characteristics. These primarily include operating records such as opening time, registration type, and lagged performances. We observe 138 primary industries, with the largest firm in each getting an average of 9.63% market share and the top 20 firms getting on average 31.6% market share. For our main analyses, we focus on several consumer electronics industries for simplicity.[5]

1.2 Data Analytics Product

Our empirical investigation of stores' information acquisition relies on a proprietary data analytics product developed and offered by the platform. Since 2015, the platform introduced a comprehensive data analytics product for stores operating on it. It includes information on aggregate demand, traffic sources, own customer demographics, and competitor strategies. Most of such information is proprietary to the platform and divided into several modules as shown in the Table 2.

[5] The industries in our main dataset include "flashdrive and storage", "DIY computer", "computer hardware, peripherals, and monitors", "electronics components", "3C electronics accessories (Consumer electronics, Communications, and Computers)", "digital camera", "kitchen electronic appliances", "AI equipment", "MP3/MP4/iPod/recorder", "network and internet equipment", "home electronic devices", "video and audio electronics", "general electronics and electrical devices".

Table 2: Modules Included in the Data Analytics Tool

Module	*Information Provided*	*Price/yr (RMB)*
Basic	Basic store-level information	Free
Traffic	Detailed traffic sources, conversion analysis, and customer demographics	888
Market Std.	Aggregate demand and trends top keywords for SEO	900 - 1188

Note: Prices vary only across time. The competition module is only available standalone for a subset of time periods. It is then bundled into the market pro module.

The traffic module provides detailed analyses of the composition of visitors (traffic) and customers that the store attracts or have attracted in the past. stores can segment their customers by several observable characteristics and understand how they discovered the store and their browsing activities within the store and across many products.

The market modules includes information of the stores' respective industries. Stores' industry affiliation is defined by the platform based on minimum sales and product offering requirements in each industry. A C store hoping to purchase the market modules in a particular industry must first accumulate some sales and offer at least one product in that industry for certain period of time. The standard market module shows trends in aggregate demand both in terms of realized transactions and in terms of top keywords searched. The trends are shown in real time and for historical periods up to one year. The pro market module extends the historical records to three years and includes additional information on the top-selling and top-searched stores, brands, and products.

The competition module identifies stores' competitors based on realized traffic flows between stores. stores can also self-define a small number of stores as competitors. After identifying competitors, the module allows stores to monitor their competitors' overall performance and the strategies used by them. Specifically, stores can monitoring whether their competitors changed product titles, listed new products, changed prices, and altered keywords bid.

Once a store activates the basic version, it has access to the data analytics tool's web and mobile interfaces. Paid modules appear after the store successfully purchase them, and examples of their content are shown in the Figures A.1, A.2, and A.3 in the appendix.

2 Reduced-Form Evidence

In this section, we document four sets of descriptive and reduced-form facts. Our goal is to make four arguments that help us learn about firms' decision to acquire information.

2.1 Data Analytics Adoption

Fact 1: Prevalence of information acquisition grows gradually but concentrates among large stores and incumbents We first describe the acquisition pattern of different data modules across stores size groups. In Figure 1, we use an area plot to demonstrate the composition of stores who use and who pay for the data analytics product, which corresponds to the axis on the left. The dark blue areas indicate the portion of stores, at a given time, that do not even activate the basic (free) version of the data analytics product. The firm uses this basic version to communicate some crucial business statistics to stores and to introduce the data analytics tool. It measures store-level traffic, sales, and transaction statistics.

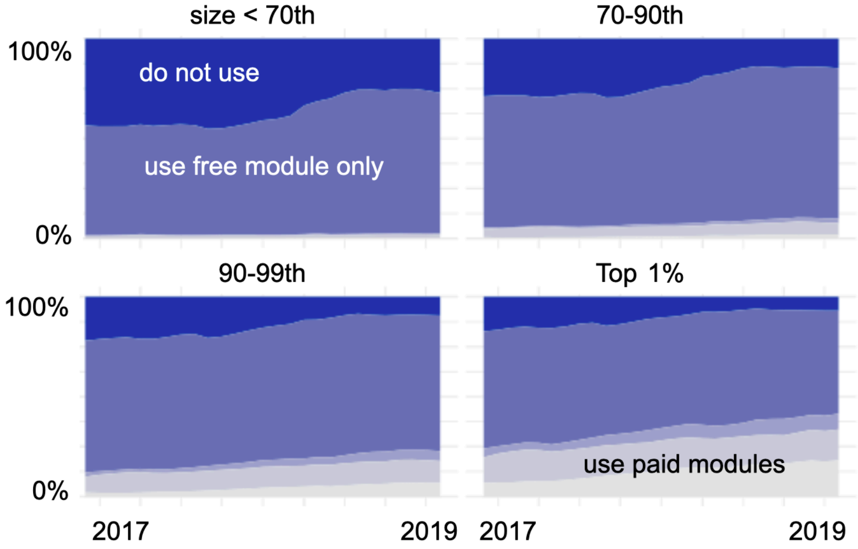


Figure 1: Adoption Patterns across Modules, Size Groups, and Time

Notes: adoption patterns of C stores by relative scale over time. Adoption is measured by activation of the corresponding data module. The four quadrant represent store size group, which is measured based on sales in the previous month, those from 0 to 40th percentile are micro stores and are omitted; 40th to 70th percentile are small stores; 70th to 90th percentile are medium stores; 90th to 99th percentile are large stores; those in the top 1 percent are very large stores.

First notice that the basic version of the data analytics product has been adopted by a majority of firms operating on the platform. This demonstrates a growing awareness of data acquisition and the type of information available through the data analytics product. However, the acquisition of costly information remain rare in comparison. Only among largest stores (90th percentile in last-month-sales) is the acquisition rate for paid modules above 10%: lighter blue colors in Figure 1 correspond to the adoption of different types of costly information. Nonetheless, based on changes in the popularity of the paid modules (information on detailed traffic breakdown for own store and that on the stores' respective market), stores of all sizes seem to have become increasingly likely to pay for data over time.

Looking at each type of costly information separately, as shown in Figure 2, the adoption patterns seem to be consistent: linear growth in adoption rates (blue line, right axis) with concentration among large incumbents (red line, left axis). In fact, the revenue share of stores that purchase costly information is around 10 to 30 times the raw fraction of those stores on the platform. It also seems that stores are particularly interested in obtaining market information, although a proper evaluation of stores' valuation of such information must account for price differentials and store heterogeneity, which we turn to next.

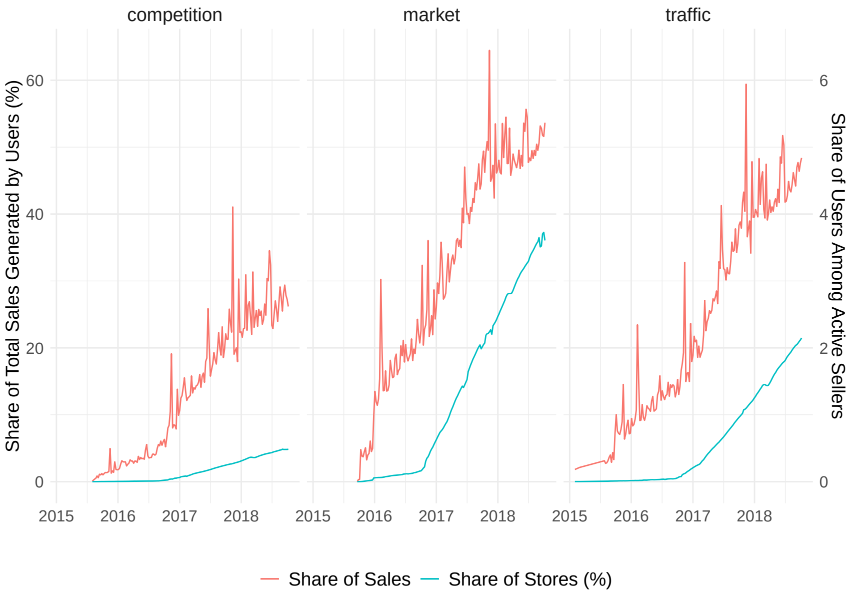


Figure 2: Adoption Patterns across Modules Over Time

Notes: the graph shows adoption patterns over time for a random sample of B and C stores lumped together. Share calculated over active stores at particular point in time.

2.2 Store Sensitivity to Cost of Information

Fact 2: stores, especially small ones, are very sensitive to the cost of acquiring information We now take advantage of a sharp price change in the market module to understand stores' sensitivity to the cost of the information modules.

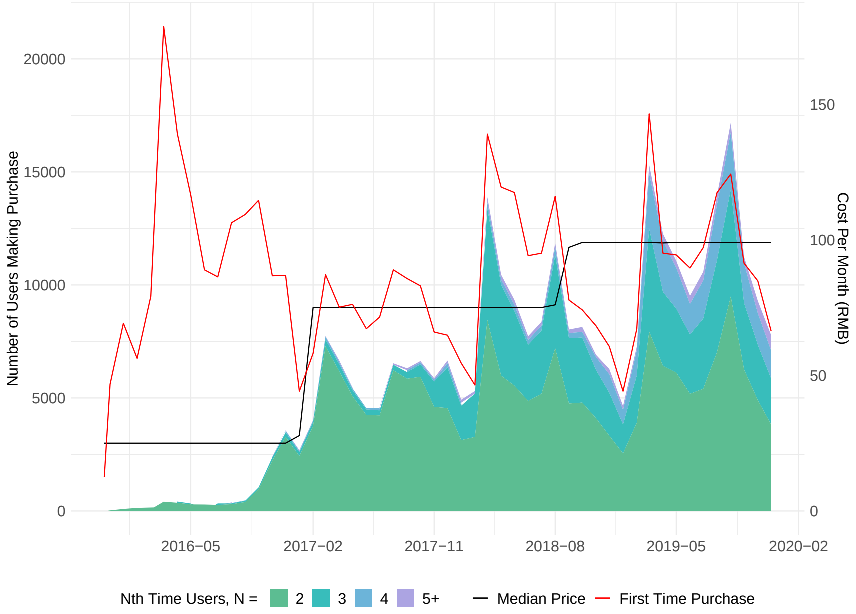


Figure 3: Adoption and Price Change of Market (Std) Module

Notes: number of orders placed for the Market Standard module. Red line depicts number of orders placed by first time users, while the colored area graph differentiate renewal purchases based on the number of renewal. Number of users are re-scaled.

The black line in Figure 3 plots the pricing of the standard market module offered in the data analytics tool, which corresponds to the right axis. The red line plots the amount of first time purchase of the corresponding module, which corresponds to the left axis. Figure A.4 in appendix shows similar trend for adoption of the Market (Pro) module. There is obvious seasonality in consumers' purchase decision. But taken that into account, it seems that quantity falls in response to each price hike. The same pattern is true for stores' renewal decision, which is represented by the colored area graphs based on number of renewal.

On top of the market information provided in the standard market module, stores can acquire additional information on the competition of top stores, brands, and products by paying more for the market pro module. The pattern and timing of price change is similar.

We now use an instrumental variable approach to make use of the price hikes and quantify stores' price sensitivity. Define instrument z for price p as follows.

$$z_{it} = \begin{cases} 1 & \text{if t >= price change date} \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

The main specification is simply

$$\ln Price_{it} = \delta z_{it} + x'_{it}\beta + \epsilon_{it} \tag{2}$$

$$Purchase_{\text{mkt},it} = \theta \ln \hat{Price}_{it} + x'_{it}\alpha + \varepsilon_{it} \tag{3}$$

Results are presented in Table 3 below. For the Market (standard) module (columns 4-6) that contains aggregate demand data for transactions and for search keywords, a 100 RMB increase in price would see adoption drop by $\sim 35\%$.[6] This number is further amplified among small or young stores, reaching 55% for those that are below 40th (B stores) or below 70th (C stores) percentile measured by previous period's sales. The same pattern can be found for the pro version, where the demand elasticity is measured at -0.0004, and at -0.0006 among small stores.

[6] This is calculated based on the pre-period price 900RMB and the pre-period average adoption rate of 0.05%.

	Dependent variable:					
	$\mathbf{1}_{std}$ (agg. demand/consumer char.)			$\mathbf{1}_{pro}$ (+ top seller/product/char)		
Price	-0.002***	-0.002***	-0.002***	-0.0004***	-0.0004***	-0.0004***
	(0.0001)	(0.0001)	(0.0001)	(0.00004)	(0.00004)	(0.00004)
Price × Small Seller		-0.001***			-0.0002***	
		(0.00003)			(0.00001)	
Price × New Sellers			-0.0001**			-0.00001
			(0.00003)			(0.00001)
Stores FE	Y	Y	Y	Y	Y	Y
Year FE	Y	Y	Y	Y	Y	Y
Month FE	Y	Y	Y	Y	Y	Y
R^2	0.107	0.107	0.107	0.047	0.047	0.047
Adjusted R^2	0.037	0.038	0.037	0.099	0.099	0.099

Table 3: Demand Elasticity of Market Modules

Notes: Regressions based on a random sample of about 52k stores in industries related to consumer electronics from 2017 to 2018. A store is in the sample if she is not a current user of the corresponding modules. Prices for a 12-month subscription are in log. Instrument is a dummy variable that equals to 1 if t is after the price change date and 0 otherwise. All regressions include stores, year and week of year fixed effects as well as variables indicating current renewal cycle by module. Regressions with interactions use z_{it} and lagged indicator for being small/young as first stage instrument for p_{it} and p_{it} interacting with current period indicator. A store is small if the current period sales is below the 70th percentile for C stores and below 40th percentile for B stores. A store is young if registration date is within a year. ***p<0.01.

There is important caveat to our analysis in this section. First, at the same time as the second price hike, the Market (pro) module also became more comprehensive as it swallowed contents from the previous competition module. Meanwhile, the price hike was announced a few days before the implementation, which spurred advanced purchase of data analytics modules. Therefore, our estimates here are only indicative of the actual demand elasticity for information. Nevertheless, our results demonstrate a general sensitivity to price and sharp difference between large and small stores. We propose a structural model in the next section to account for changes in bundling=.

2.3 Data Acquisition and Strategy

Fact 3: Information acquisition is tied to product and marketing strategies What do firms acquire costly information for? In this section, we conduct several tests to link stores' strategies to their decision to acquire information.

To do so, we first look at the timing of actual visiting patterns of different types of information and that of strategy implementation. We conduct a simple regression where the dependent variables are strategies implemented, while the independent variables are visit histories of the different information modules during the same week. Specifically, let i index stores and j the information modules, we can conduct a horse-race regression as follows.

$$Strategy_{it} = \alpha_i + \alpha_t + \sum_{j \in J} \theta_j Visit_{ijt} + x'_{it}\beta + \epsilon_{ij}$$

The estimated θ_j's for various modules j is presented in the graph below, where different color represents different regressions based on the dependent variable, product strategies.

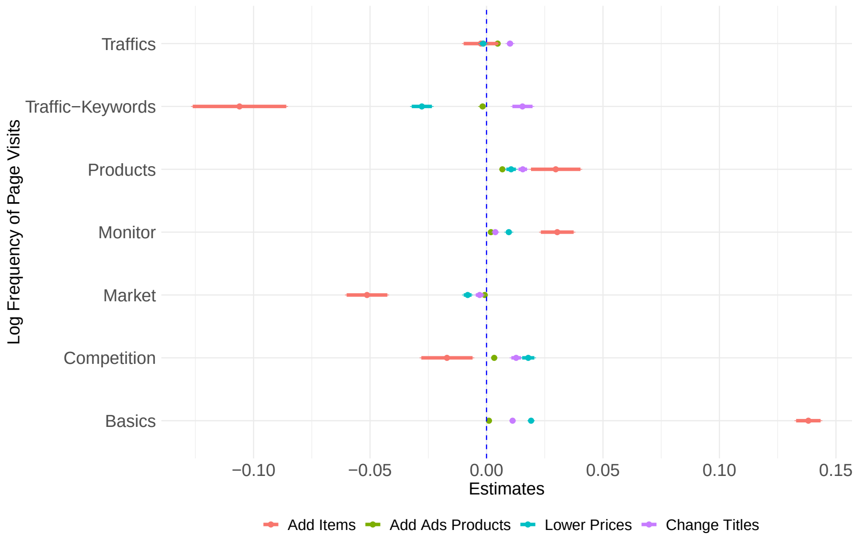


Figure 4: Strategy Implementation and Information Acquired

Notes: visit patterns calculated over a sample of stores in consumer electronics related industries from March to August 2018 (prior to price changes and bundling of market modules). All strategies are measured in frequencies and are in log. Visit patterns are captured by dummies indicating having visited any pages in the corresponding module. Additional control variables include number of items listed, current week's total sales and traffics. All regressions also include store and week of year fixed effect.

Lastly, we conduct an event study around the acquisition of different data modules to understand the dynamics of different strategies a firm implements when it decides to acquire information.

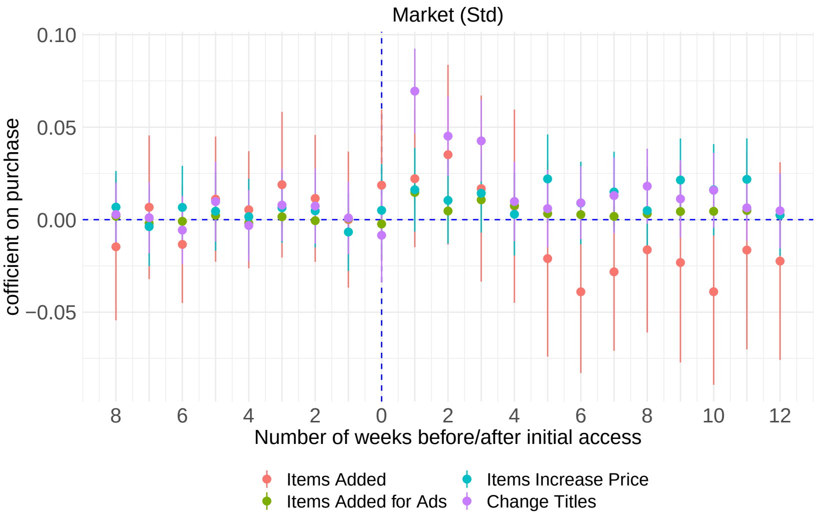


Figure 5: Strategy and Timing of Adoption: Market Info

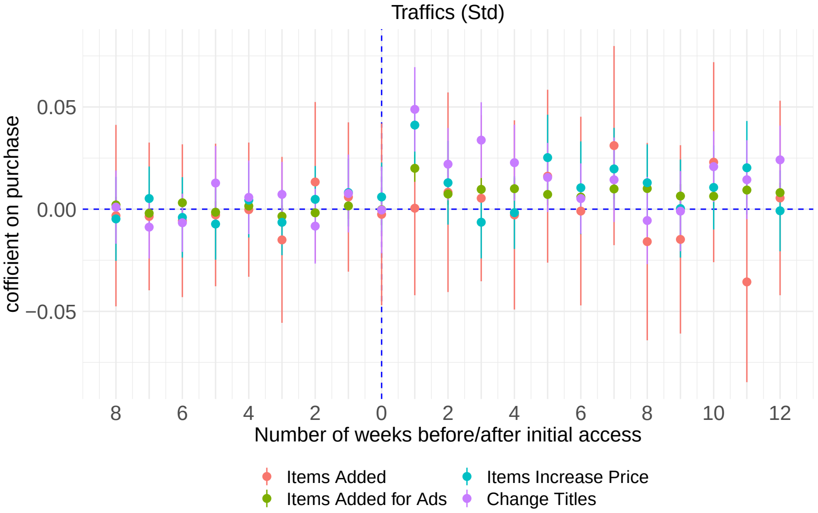


Figure 6: Strategy and Timing of Adoption: Own Traffic Info

*Notes:*sample consists of stores in consumer electronics related industries that have ever purchased market module. Regressions include store and time fixed effects. Each dot-bar represents coefficients on a dummy equals 1 if for store i current week t is j weeks before/after subscription to market module start. Omitted category is 13 or more weeks before start. Dependent variables of frequency of employing certain product strategies measured in log or log average price paid per buyer.

Figures 5 and 6 demonstrate the frequency of product strategies implemented with a fixed effect regression, where we take all stores who have purchased the market module, and regress strategies implemented on time relative to adoption (i.e. a dummy equal 1 if a particular week t is k week before/after adoption for store i) and store as well as calender week fixed effects.

$$Strategy_{it} = \alpha_i + \alpha_t + \sum_{j \in \{-12, \ldots 12\}} \theta_j RelativeWeek_{ijt} + RenewalCycle_{it}\beta + \epsilon_{ij}$$

The results presented in Figures 5 and 6 confirm a correlation between the implementation of product strategies and the acquisition of costly information on market and own traffic. Our sample does not allow us to conduct the same test for the competition modules and the information contained within due to the low takeup rate. On the other hand, price levels seem to have dropped slightly after stores acquire data on their stores' own traffic breakdown, but the magnitude of the effect appear small and become undetectable after a few weeks.

2.4 Impact of Data Acquisition

Fact 4: Information acquisition facilitate growth Does acquiring information actually make implementing strategies more effective, facilitating revenue growth?

An easy way to get at this effect is to use a matching estimator and look at the effect of information acquisition on revenue growth, which is a particularly important metric for small- and medium-sized businesses. To do so, we first construct a matching sample for the group of stores that have purchased each data module. Specifically, we match on all observable characteristics of the store, including size, age, and industry dummies, etc. We also match stores based on their past sales and strategy performances. Please find details of the matching estimator in Appendix C. Importantly, we only match stores based on their statistics more than 7 weeks prior to the acquisition time. This allows us to visualize pre-trends that may indicate unobserved differences between the adopting store and the matched control stores.

We can then conduct the following regression. Here, X includes variables used in the matching function for all eleven weeks prior to t_0: observable characteristics of the store, actions, and performance. This means that the outcome variable at time t_0 is included in X.

$$\ln Sales_{i,t_0+k} = \gamma_k Adopt_{i,t_0} + X'_{i,t_0}\beta + \alpha_k + \epsilon_{k,i} \; \forall k = 1, ..., 11$$

Figure 7 plots the implied treatment effect γ_k across time k on the horizontal axis, with different colors representing different type of information acquired. We are able to detect sizable treatment effect based on comparing stores that have identical observable characteristics as well as performance and strategy profiles. Results from matching estimator suggest that stores that paid for the information indeed out-performed stores that do not, especially the modules being purchased are about markets and traffic. Moreover, as shown in Figures 8 A.5 and A.6, comparing to large stores, small stores are able to achieve higher growth (in log terms) relative to their matched counterparts.

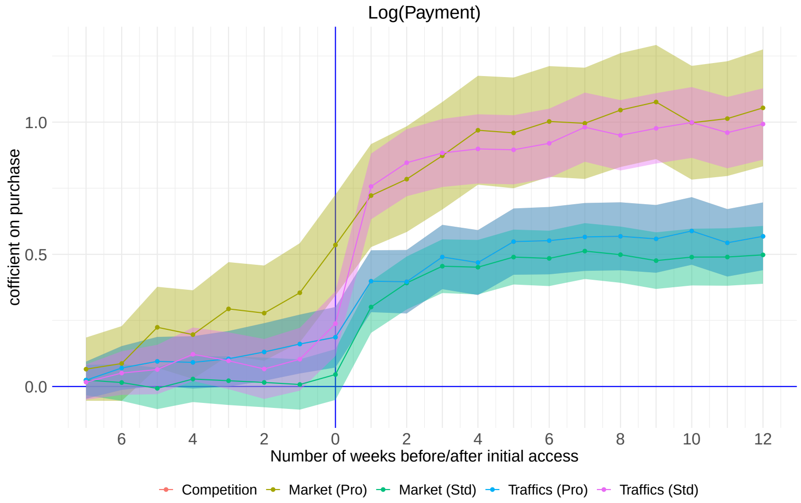


Figure 7: Effects of Data Acquisition on Log Sales

Notes: this graph shows coefficient γ_k for $k = -11, ..., 11$ where $k = 0$ is the week of adoption, separately for each information module. Sample size for each information varies and is determined by number of users in the consumer electronics industry sample that have adopted corresponding module for the first time between 2017 to mid 2018. Each matched sample is constructed independently following procedure described in the appendix. Dependent variable is log sales in the previous week. All regressions include controls for stores characteristics, performance and actions profile in the pre-adoption period.

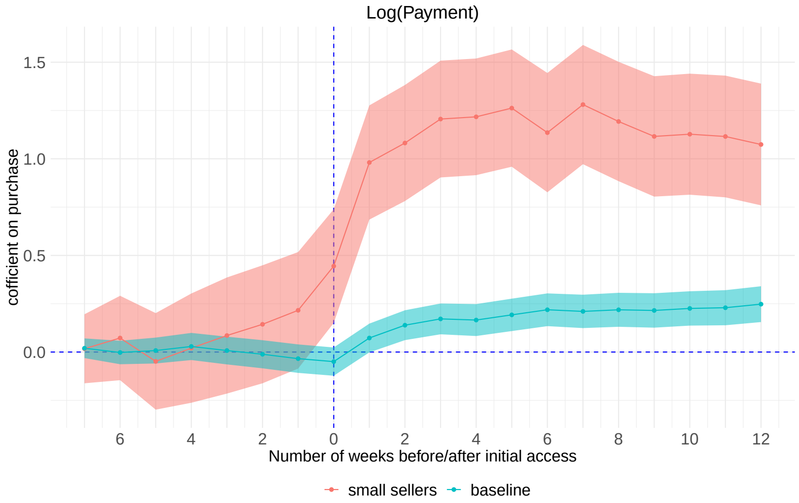


Figure 8: Effects of Acquiring Aggregate Demand Information on Log Sales, by store Size

Notes: this graph repeats Figure 7 for the standard market only, but by store size groups.

The key weakness of our approach is that there may exist unobserved motivation/skill/cost differences across the matched and matching samples that endogenously determine information acquisition but are not captured by pre-period actions and outcome. Meanwhile, due to small number of stores that actually purchased competition module within our sample, the test might be under-powered.

3 A Model of Information Acquisition and Product Strategy

Following the evidence presented in the section above, we propose a model of information acquisition and product strategy. In order to understand firms' decision to acquire information together with their desire to implement product strategies, we must estimate two types of key parameters. First, stores' valuation of different types of information in relations to each product strategy (γ). Second, stores' costs of implementing product strategies (c).

Let i index stores, j DAP (data analytics product) module (bundles), t time period (week), and k information that can be bundled into DAP modules. Further denote the characteristics of a store as x and its renewal count as rc. On the other hand, denote stores' product strategy as s. Our focus here are the number of times new products are introduced or existing product re-positioned through title changes.

The most important primitives of the model is stores' valuation of how different pieces of information can lead to increasing effectiveness in executing product strategies, as reflected by bringing in additional sales growth. We model these valuations and correlations directly. For store i at time t, conditional on the stream of product strategies that it would implement, $\mathbf{s}_{it}$, its valuation of each DAP module j would be as follows.

$$v_{ijt}(\mathbf{s}_{it}) = \mathbb{E}\left[\sum_{s}\sum_{k\in K_j, t'\in T_j} \gamma_{s,ikt'} \cdot s_{it'}\right] \tag{4}$$

$$\text{where } \gamma_{s,ikt} = \gamma_s(x^{\gamma}_{it}, \mathbf{1}_{yr}, \mathbf{1}_{week}, \mathbf{1}_k; \theta_\gamma) + \epsilon_{s,ikt} \tag{5}$$

$$\epsilon_{s,ikt} \sim \mathbf{N}(\mathbf{0}, \mathbf{\Sigma}_\gamma(\sigma_\gamma, \rho_{\gamma,t}, \Sigma_k))] \tag{6}$$

Here, x^γ represents observable characteristics of the store, including firm age, size, business type, and industry dummies. 1_{week}, 1_{yr}, and 1_k are fixed effect for week of year, year, and content. Lastly, as in Equation 6, γ is allowed to vary based on an unobservable component ϵ_s. It is assumed to be i.i.d. across stores i and follows a joint normal distribution Σ_γ. Σ_γ is parameterized so that ϵ_s are correlated across content types k based on Σ_k, and across time based with serial correlation term $\rho_{\gamma,t}$. Notice that, in this model, we assume that γ_s's are independent across different sets of strategies.

Based on the effectiveness of acquiring information, we can specify a simple logit models to understand stores' choices to acquire information and to implement product strategies.

$$u_{ijt}(\mathbf{s}_{it}) = v_{ijt}(\mathbf{s}_{it}|\gamma) + \alpha_{it}p_{ijt} + \xi_j + \varepsilon_{ijt} \tag{7}$$

$$\text{where } \alpha_{it} = \alpha_0 + \alpha_1 x_{it}^{\alpha} \tag{8}$$

$$\text{and } \varepsilon_{ijt} \sim EV1(0, \frac{\pi^2}{6}) \tag{9}$$

$$\text{so that } \Pr(\mathbf{1}_{ijt}|\mathbf{s}_{it}) = \Pr\left(u_{ijt}(\mathbf{s}_{it}) = \max_{j'}\left[u_{ij't}(\mathbf{s}_{it})\right]\right) \tag{10}$$

Here, p_{ijt} denotes the price for the information module j. stores' information purchase decisions are intertwined with their desire to implement product strategies s, taking into account their valuation of the complementarity between information acquisition and strategy implementation, γ. Let $\mathbf{s}_{it} = (s_{it}, ..., s_{i,t+12})$ and each s_{it} be a binary indicator for strategy implementation.

$$\Pr(s_{it} = 1|\{\mathbf{1}_{ijt}\}_J) = \Lambda\left[\sum_j \mathbf{1}_{ijt} \cdot v_{ijt}(s_{it}|\gamma) + c_{s,it} \geq 0\right] \tag{11}$$

$$\text{where } c_{s,it} = x_{it}^{c\prime}\theta_c \tag{12}$$

Similar to Equations 5 and 6, we specify c as a random effect that denotes the unobserved baseline net gain of implementing strategies s. We assume that c_s's are independent across different sets of strategies and that they are only correlated with γ_s's based on observables. Notice that we do not consider the cost of information acquisition p here. This is because that cost would be sunk once the store has made the acquisition decision, therefore, any subsequent strategy decision should not depend on it.

Lastly, assume that store revenue follows a log-normal process in which the mean depends on observable characteristics and lagged sales, as well as the expected gain from implementing strategies. For our baseline version, we assume that stores are, on average, rational in their expectation of the impact of information acquisition.

$$r_{it} \sim N(\mu_{r,it;\theta_r}, \sigma_r) \tag{13}$$

$$\text{where } \mu_{r,it} = x_{it}^{r\prime}\theta_r + \sum_j \mathbf{1}_{ijt} \cdot v_{ijt}(\mathbf{s}_{it}|\gamma) \tag{14}$$

For simplicity, we start with one $\mathbf{s}'_{it}$, product title change, and three type of information, aggregate demand, top stores/product/brands, and competition. This reduces the computational burden.

4 Model Estimation, Results, and Counterfactuals

4.1 Model Estimation

We estimate the model with a simulated maximum likelihood approach focusing on the likelihood of streams of (1) product strategies implementation: adding product and changing product titles; (2) acquisition of market and competition information; and (3) sales outcome.

For each set of parameter proposal, we generate 50 pseudo-random draws of $\epsilon_{s,ijt}$ and compute the likelihoods. Averaging over $\epsilon_{s,ijt}$ for the corresponding module j gives us a set of v_{ijt}'s, which we can directly use to calculate likelihoods in equations 10, 11, and 13.

4.2 Identification

We estimate the model on a subsample of 500 stores and 18 periods are reported in the following section, while bootstrapping for standard error. The time subset is chosen to include both pre- and post-price change in our instrumental variable regression in Equation 1. During the price change, competition information is merged into the market pro module, while the competition module becomes unavailable.

First, we can compare sales outcome and information acquisition decisions both across stores and within stores, conditional on observables and lagged performance. Our setting also allows us to do so under different pricing and contract space environments. This helps us pin down the scale of γ's and α's, conditional on the observed strategy streams $\mathbf{s}_{it}$. How γ's change over time and correlates across different content types k allows us to fit the linear model in equation 5 and obtain θ_γ, σ_γ, $\rho_{\gamma,t}$, and Σ_k.

Conditional on the effect of acquiring data and the realized information acquisition decisions ($\mathbf{1}_{ijt}$), we can then maximize the likelihood of observing the strategy stream for each store. Ideally, we would want to observe situations where information acquisition would not be possible, therefore deducing the likelihood that acquiring information would make a difference. But given that the v's are already identified above, the observed frequency of strategy alone is enough to pin down the levels of c and therefore the θ_c's

4.3 Estimation Results

Estimation results are presented in Tables B.1.

Our price sensitivity estimates capture the same pattern across store size groups as in the reduced-form section above. Log store size, as measured by lag revenue, strongly and negatively contributes to the magnitude of overall price sensitivity. Conditional on store size, however, older stores are actually more price sensitive than newer ones. On the other hand, γ, the benefit of acquiring market information is higher for older stores, while newer stores benefit more from acquiring competitor information. ξ terms capture the premium on each module or information piece unexplained by revenue impact (γ), relative to the market pro module and the top rankings information. The expensive market pro module seem to command a premium over its cheaper counterparts, which can explain why its purchase share is not small in comparison with the other two modules, despite significantly higher prices. However, the content contained in that module, the top rankings (of stores, products, and brands), actually have inferior baseline revenue impact.

Using these hyper-parameters, we can calculate latent parameter distributions, reported in Table 4. There is large variations in price sensitivity, as identified by differential adoption responsiveness to cost changes in data acquisition over time across store size and age groups.

In light of the magnitude of α's, the estimates of γ parameters indicate that the effect of information acquisition, which is estimated based on actual sales performance, can be very large. However, there are significant heterogeneity across stores.

Table 4: Latent Parameter Distributions

Statistic	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
α(size,time)(e-4)	-9.09	5.6	-21.97	-14.86	-7.84	-4.78	1.18
$\gamma_{\text{top rankings}}$	0.42	2.36	-5.07	-1.27	0.40	1.96	20.41
$\gamma_{\text{agg demand}}$	1.66	1.62	-13.72	0.59	1.60	2.68	6.33
$\gamma_{\text{competitors}}$	2.51	2.14	-5.32	1.01	2.69	4.17	7.57

We can further dissect the sample of stores by size and age to examine the heterogeneity in the impact of data on sales. Figure 9 shows our results with dissection across firm size. It demonstrates that although the median firm in both size groups both have positive valuation for all types of information, there exists significant heterogeneity within each size group. Meanwhile, comparing median differences across size groups, supply (competitor) information and traffic details (own customer sources and demographics) seem to be particularly useful among small stores. Only aggregate demand information is more highly-valued among larger stores.

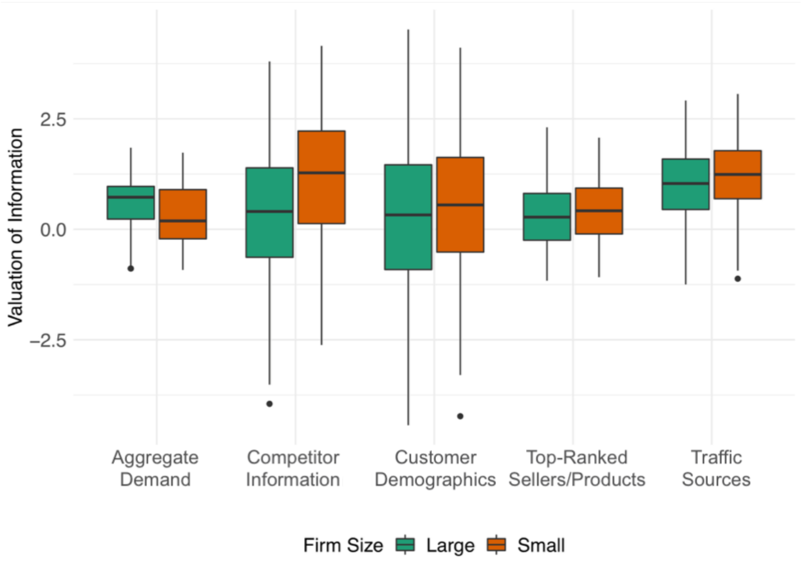


Figure 9: Effects of Information Acquisition on Log Sales across Data Type and Size Groups

Notes: this is a box plot of the value of information, measured in $\hat{\gamma}$ estimate from our model across information type and across store size group (50-50 dissection).

Figure 10 repeats the above analysis, but dissecting stores based on age. It confirms and amplifies three patterns that we have seen above in the comparison across firm size (firm size and age are highly correlated). Aggregate demand information is very valuable for incumbents, while being virtually useless for the average young firm.

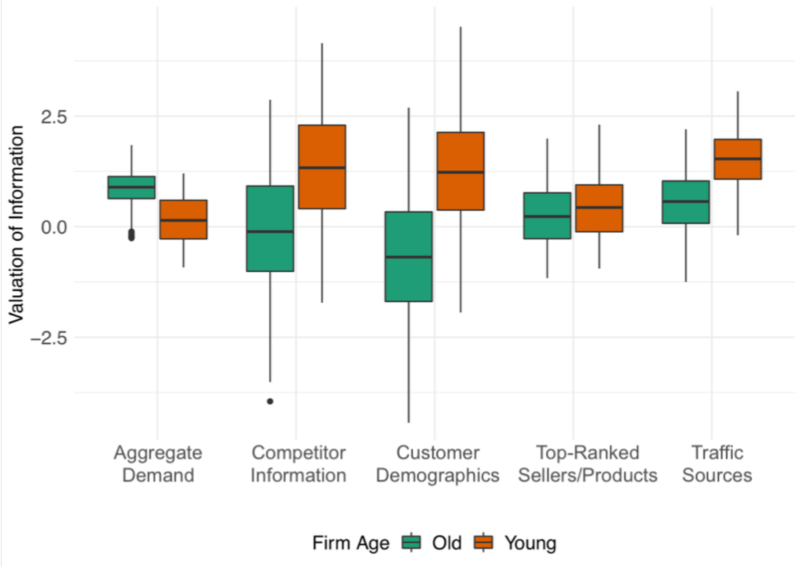


Figure 10: Effects of Information Acquisition on Log Sales across Data Type and Age Groups

Notes: this is a box plot of the value of information, measured in $\hat{\gamma}$ estimate from our model across information type and across store size group (50-50 dissection).

4.4 Counterfactual Data Pricing: Marginal Changes in Overall Growth and Concentration

With the demand estimates, we can conduct simple counterfactual simulations that alter the pricing of the data product. The outcome that we focus on is the marginal impact on overall growth in sales across all stores and that on market concentration, as measured by the Herfindahl-Hirschman Index (HHI) among the stores in our sample.

Figure 11 plots the simulated outcomes, where the color gradient of each dot represents the intensity of counterfactual discounts applied uniformly across all data modules and for all stores. Darker means deeper discount. On the vertical axis, we show marginal changes in overall sales growth across all stores, while the horizontal axis shows the marginal change in market concentration.

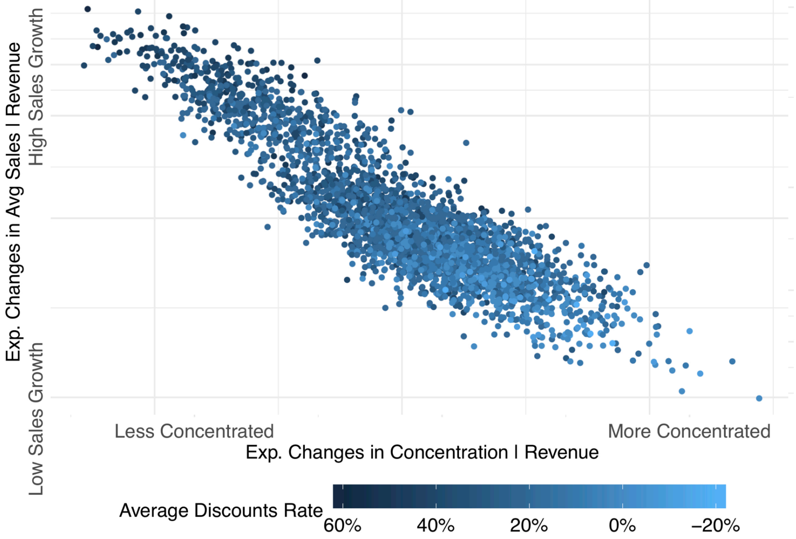


Figure 11: Effects of Data Acquisition on Log Sales

Notes: this graph shows the marginal impact on overall sales across all stores (vertical axis), and on industry concentration (horizontal axis) as measured by the HHI index when the platform applies uniform discounts on the all data an analytics products.

Raising the cost of data depresses overall sales growth, and it does so by disproportionally suppressing data acquisition among small stores, therefore increasing concentration. Sales growth among large firms bring in immediate revenue for the platform, while sales growth among small stores may beget larger gains in future growth of the platform. The platform can therefore balance these forces to determine the profit-maximizing prices on different types of data.

5 Conclusion

In this paper, we acquire detailed proprietary data about online stores on a large e-commerce platform. They allow us to uncover several key facts about how different firms value different types of data, the strategies that are influenced by those data, as well as the ultimate growth impact.

The implication of our work goes beyond the online retail setting. When information is costly as opposed to freely available in the market (credit history data of borrowers can be easily accessed by almost all accredited lenders and insurers in the US, for example), our results imply that large firms are much more likely to acquire information than their smaller counterparts. We also demonstrate that information acquisition *can* lead to significant sales growth, often through non-price channels, which may exacerbate market concentration. To mitigate this endogenous force towards concentration, a social planner can reduce the cost of information, particularly among small and young firms and for information on their own customers and operations as well as on their competitors.

References

Arcidiacono, Peter, Paul B Ellickson, Carl F Mela, and John D Singleton (2016). "The Competitive Effects of Entry: Evidence from Supercenter Expansion". In: *Working Paper, Duke University*.

Armstrong, Mark and Steffen Huck (2010). "Behavioral economics as applied to firms: a primer". In:

Atkin, David, Azam Chaudhry, Shamyla Chaudry, Amit K Khandelwal, and Eric Verhoogen (2017). "Organizational barriers to technology adoption: Evidence from soccer-ball producers in Pakistan". In: *The Quarterly Journal of Economics* 132.3, pp. 1101–1164.

Bai, Jie (2018). "Melons as Lemons : Asymmetric Information , Consumer Learning and Quality Provision". In: *Working Paper, Harvard Kennedy School*.

Bajari, Patrick, Victor Chernozhukov, Ali Hortaçsu, and Junichi Suzuki (2019). "The impact of big data on firm performance: An empirical investigation". In: *AEA Papers and Proceedings*. Vol. 109, pp. 33–37.

Bergemann, Dirk and Stephen Morris (2019). "Information design: A unified perspective". In: *Journal of Economic Literature* 57.1, pp. 44–95.

Berto Villas-Boas, Sofia (2007). "Vertical Relationships between Manufacturers and Retailers: Inference with Limited Data". In: *The Review of Economic Studies* 74.2, pp. 625–652.

Blake, Thomas, Sarah Moshary, Kane Sweeney, and Steven Tadelis (2018). *Price Salience and Product Choice*. Tech. rep. National Bureau of Economic Research.

Bloom, Nicholas, Aprajit Mahajan, David McKenzie, and John Roberts (2010). "Why do firms in developing countries have low productivity?" In: *American Economic Review* 100.2, pp. 619–23.

Brynjolfsson, Erik, Lorin M Hitt, and Heekyung Hellen Kim (2011). "Strength in numbers: How does data-driven decisionmaking affect firm performance?" In: *Available at SSRN 1819486*.

Brynjolfsson, Erik and Kristina McElheran (2016). "The rapid adoption of data-driven decision-making". In: *American Economic Review* 106.5, pp. 133–39.

Cole, Shawn and Asanga Nilesh Fernando (2012). "The Value of Advice: Evidence from Mobile Phone-based Agricultural Extension". In: *Working Paper, Harvard Business School*.

Conley, Timothy G and Christopher R Udry (2010). “Learning about a new technology: Pineapple in Ghana”. In: *American Economic Review* 100.1, pp. 35–69.

DellaVigna, Stefano and Matthew Gentzkow (2017). *Uniform Pricing in U.S. Retail Chains*. Tech. rep. National Bureau of Economic Research.

Dubé, Jean-Pierre, Zheng Fang, Nathan Fong, and Xueming Luo (2017). “Competitive Price Targeting with Smartphone Coupons”. In: *Marketing Science* 36.6, pp. 944–975.

Dubé, Jean-Pierre and Sanjog Misra (2017). *Scalable price targeting*. Tech. rep. National Bureau of Economic Research.

Eichenbaum, Martin, Nir Jaimovich, and Sergio Rebelo (2011). “Reference Prices, Costs, and Nominal Rigidities”. In: *American Economic Review* 101.1, pp. 234–62.

Ellickson, Paul B and Sanjog Misra (2008). “Supermarket pricing strategies”. In: *Marketing science* 27.5, pp. 811–828.

Ellison, Glenn and Sara Fisher Ellison (2009). “Search, obfuscation, and price elasticities on the internet”. In: *Econometrica* 77.2, pp. 427–452.

Ellison, Glenn and Sara Fisher Ellison (2005). “Lessons about Markets from the Internet”. In: *Journal of Economic Perspectives* 19.2, pp. 139–158.

Foster, Andrew D and Mark R Rosenzweig (1995). “Learning by doing and learning from others: Human capital and technical change in agriculture”. In: *Journal of Political Economy* 103.6, pp. 1176–1209.

Hanna, Rema, Sendhil Mullainathan, and Joshua Schwartzstein (2014). “Learning Through Noticing: Theory and Evidence from a Field Experiment”. In: *The Quarterly Journal of Economics* 129.3, pp. 1311–1353.

Hitsch, Günter J, Ali Hortacsu, and Xiliang Lin (2017). “Prices and promotions in us retail markets: Evidence from big data”. In: *Working Paper, University of Chicago Booth School of Business*.

Hjort, Jonas and Jonas Poulsen (2019). "The Arrival of Fast Internet and Employment in Africa". In: *American Economic Review* 109.3, pp. 1032–79.

Hsieh, Chang-Tai and Peter J Klenow (2009). "Misallocation and manufacturing TFP in China and India". In: *The Quarterly Journal of Economics* 124.4, pp. 1403–1448.

Illanes, Gastón (2017). "Switching Costs in Pension Plan Choice". In: *Working Paper, Northwestern University*.

Illanes, Gaston and Sarah Moshary (2015). "Estimating the Effect of Potential Entry on Market Outcomes Using a Licensure Threshold". In: *Working Paper, University of Chicago Booth School of Business*.

Jensen, Robert (2007). "The Digital Provide: Information (Technology), Market Performance, and Welfare in the South Indian Fisheries Sector". In: *The Quarterly Journal of Economics* 122.3, pp. 879–924.

Jin, Ginger Zhe (2005). "Competition and Disclosure Incentives: an Empirical Study of HMOs". In: *RAND Journal of economics*, pp. 93–112.

Jin, Yizhou and Shoshana Vasserman (2019). "Buying Data from Consumers: The Impact of Monitoring Programs in U.S. Auto Insurance". In: *Working Paper, Harvard University*.

Li, Yang, Brett R Gordon, and Oded Netzer (2018). "An Empirical Study of National vs. Local Pricing by Chain Stores Under Competition". In: *Marketing Science* 37.5, pp. 812–837.

Libgober, Jonathan and Xiaosheng Mu (2018). "Informational Robustness in Intertemporal Pricing". In: *Available at SSRN 2892096*.

McKenzie, David and Christopher Woodruff (2015). *Business practices in small firms in developing countries*. The World Bank.

McShane, Blakeley B, Chaoqun Chen, Eric T Anderson, and Duncan I Simester (2016). "Decision Stages and Asymmetries in Regular Retail Price Pass-Through". In: *Marketing Science* 35.4, pp. 619–639.

Nair, Harikesh S, Sanjog Misra, William J Hornbuckle IV, Ranjan Mishra, and Anand Acharya (2017). "Big data and marketing analytics in gaming: Combining empirical models and field experimentation". In: *Marketing Science* 36.5, pp. 699–725.

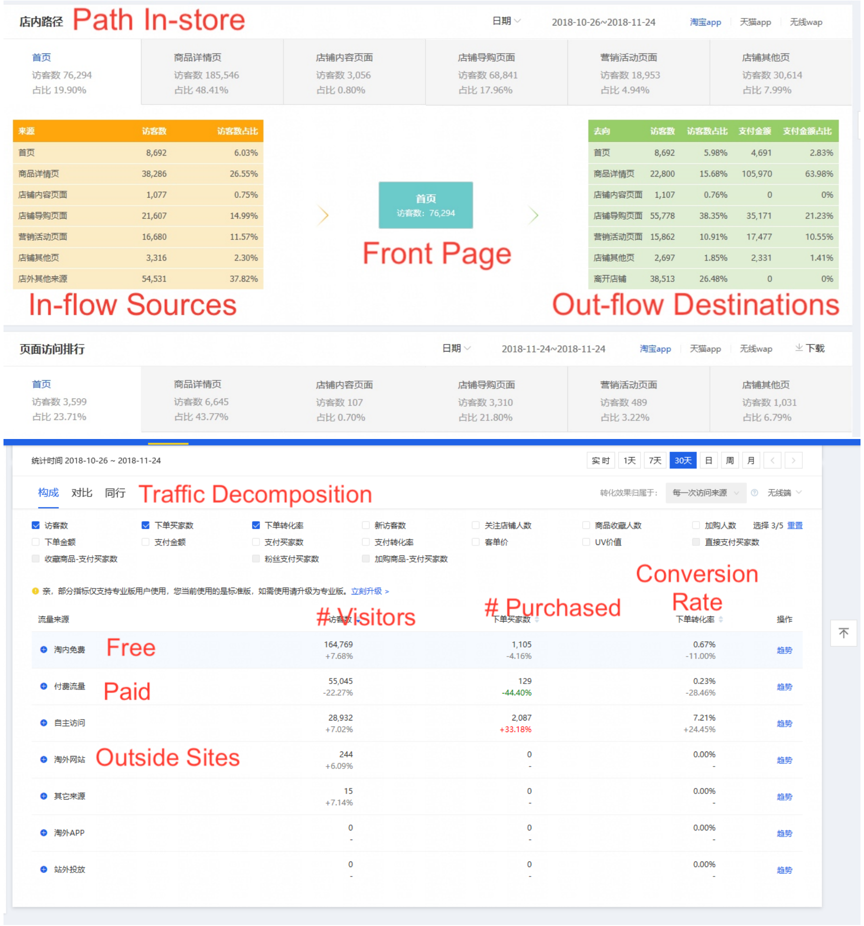
Saunders, Adam and Prasanna Tambe (2015). "Data Assets and Industry Competition: Evidence from 10-K Filings". In: *Available at SSRN 2537089*.

Tybout, James R (2000). "Manufacturing firms in developing countries: How well do they do, and why?" In: *Journal of Economic Literature* 38.1, pp. 11–44.

Wu, Lynn, Lorin M Hitt, and Bowen Lou (2017). "Data Analytics Skills, Innovation and Firm Productivity". In: *The Wharton School Research Paper* 86.

Appendix

A Additional Tables and Graphs



店内路径　日期　2018-10-26~2018-11-24　淘宝app | 天猫app | 无线wap

首页	商品详情页	店铺内容页面	店铺导购页面	营销活动页面	店铺其他页
访客数 76,294 占比 19.90%	访客数 185,546 占比 48.41%	访客数 3,056 占比 0.80%	访客数 68,841 占比 17.96%	访客数 18,953 占比 4.94%	访客数 30,614 占比 7.99%

来源	访客数	访客数占比
首页	8,692	6.03%
商品详情页	38,286	26.55%
店铺内容页面	1,077	0.75%
店铺导购页面	21,607	14.99%
营销活动页面	16,680	11.57%
店铺其他页	3,316	2.30%
店外其他来源	54,531	37.82%

首页
访客数：76,294

去向	访客数	访客数占比	支付金额	支付金额占比
首页	8,692	5.98%	4,691	2.83%
商品详情页	22,800	15.68%	105,970	63.98%
店铺内容页面	1,107	0.76%	0	0%
店铺导购页面	55,778	38.35%	35,171	21.23%
营销活动页面	15,862	10.91%	17,477	10.55%
店铺其他页	2,697	1.85%	2,331	1.41%
离开店铺	38,513	26.48%	0	0%

页面访问排行　日期　2018-11-24~2018-11-24　淘宝app | 天猫app | 无线wap　下载

首页	商品详情页	店铺内容页面	店铺导购页面	营销活动页面	店铺其他页
访客数 3,599 占比 23.71%	访客数 6,645 占比 43.77%	访客数 107 占比 0.70%	访客数 3,310 占比 21.80%	访客数 489 占比 3.22%	访客数 1,031 占比 6.79%

统计时间 2018-10-26 ~ 2018-11-24　实时　1天　7天　30天　日　周　月

构成　对比　同行　　转化效果归属于：每一次访问来源　无线端

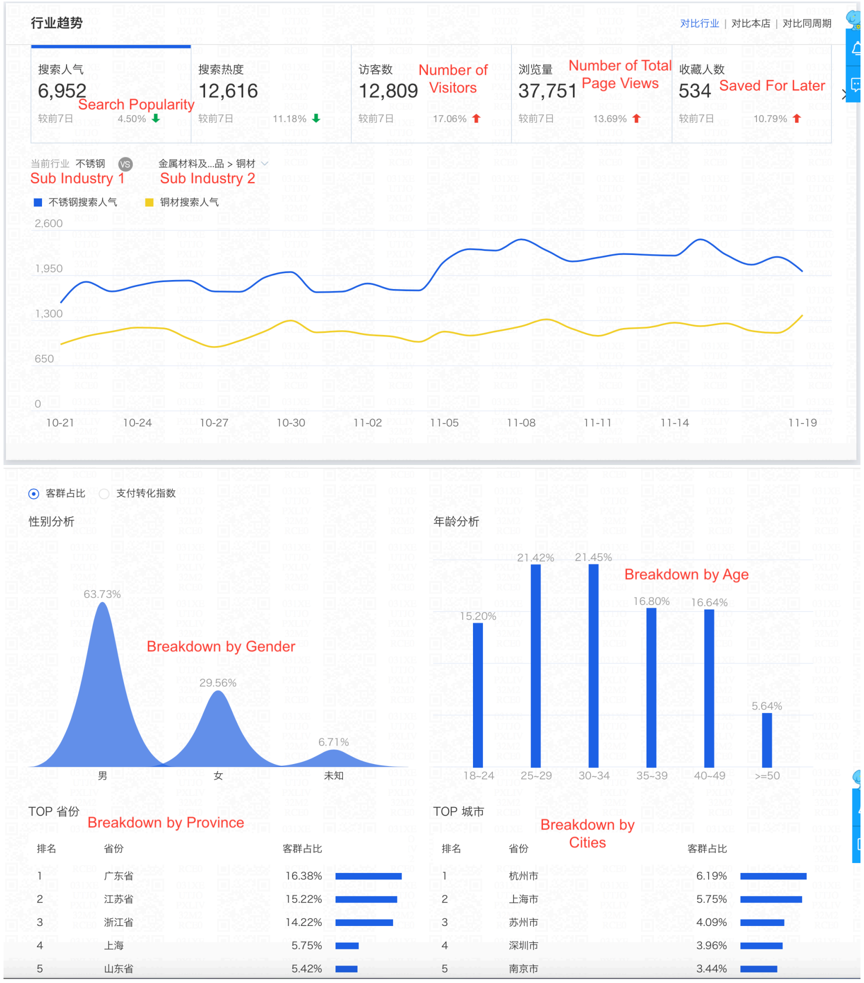
☑ 访客数　☑ 下单买家数　☑ 下单转化率　□ 新访客数　□ 关注店铺人数　□ 商品收藏人数　□ 加购人数　选择 3/5 重置
□ 下单金额　□ 支付金额　□ 支付买家数　□ 支付转化率　□ 客单价　□ UV价值　□ 直接支付买家数
□ 收藏商品-支付买家数　□ 粉丝支付买家数　□ 加购商品-支付买家数

亲，部分指标仅支持专业版用户使用，您当前使用的是标准版，如需使用请升级为专业版。立刻升级 >

流量来源	访客数	下单买家数	下单转化率	操作
淘内免费	164,769 +7.68%	1,105 -4.16%	0.67% -11.00%	趋势
付费流量	55,045 -22.27%	129 -44.40%	0.23% -28.46%	趋势
自主访问	28,932 +7.02%	2,087 +33.18%	7.21% +24.45%	趋势
淘外网站	244 +6.09%	0 -	0.00% -	趋势
其它来源	15 +7.14%	0 -	0.00% -	趋势
淘外APP	0 -	0 -	0.00% -	趋势
站外投放	0 -	0 -	0.00% -	趋势

Figure A.1: Traffic Module Interface

Notes: This is a screenshot that illustrates the web interfaces of the paid traffic module in the data analytics tool.



行业趋势

对比行业 | 对比本店 | 对比同周期

搜索人气	搜索热度	访客数	浏览量	收藏人数
6,952	12,616	12,809	37,751	534
较前7日 4.50% ⬇	较前7日 11.18% ⬇	较前7日 17.06% ⬆	较前7日 13.69% ⬆	较前7日 10.79% ⬆

当前行业 不锈钢 VS 金属材料及...品 > 铜材

○ 客群占比 ○ 支付转化指数

TOP 省份

Breakdown by Province

排名	省份	客群占比
1	广东省	16.38%
2	江苏省	15.22%
3	浙江省	14.22%
4	上海	5.75%
5	山东省	5.42%

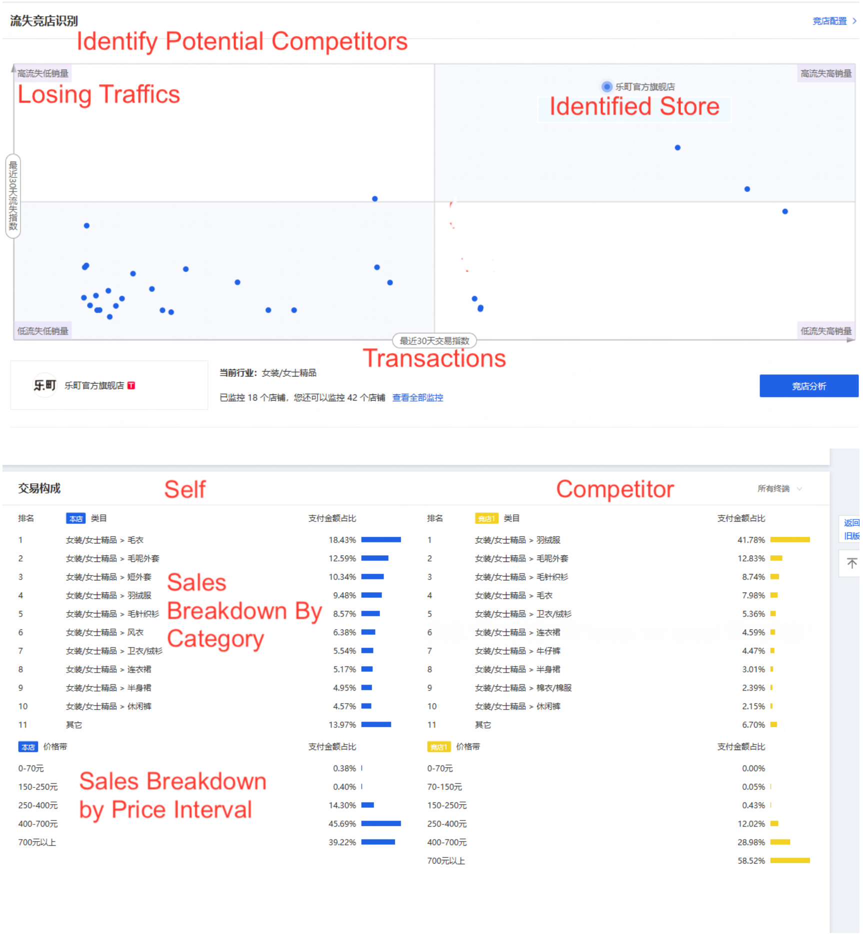
TOP 城市

Breakdown by Cities

排名	省份	客群占比
1	杭州市	6.19%
2	上海市	5.75%
3	苏州市	4.09%
4	深圳市	3.96%
5	南京市	3.44%

Figure A.2: Traffic Module Interface

Notes: This is a screenshot that illustrates the web interfaces of the paid market modules (standard and pro) in the data analytics tool.



当前行业：女装/女士精品

已监控 18 个店铺，您还可以监控 42 个店铺 查看全部监控

排名	本店 类目	支付金额占比
1	女装/女士精品 > 毛衣	18.43%
2	女装/女士精品 > 毛呢外套	12.59%
3	女装/女士精品 > 短外套	10.34%
4	女装/女士精品 > 羽绒服	9.48%
5	女装/女士精品 > 毛针织衫	8.57%
6	女装/女士精品 > 风衣	6.38%
7	女装/女士精品 > 卫衣/绒衫	5.54%
8	女装/女士精品 > 连衣裙	5.17%
9	女装/女士精品 > 半身裙	4.95%
10	女装/女士精品 > 休闲裤	4.57%
11	其它	13.97%

本店 价格带	支付金额占比
0-70元	0.38%
150-250元	0.40%
250-400元	14.30%
400-700元	45.69%
700元以上	39.22%

排名	竞店1 类目	支付金额占比
1	女装/女士精品 > 羽绒服	41.78%
2	女装/女士精品 > 毛呢外套	12.83%
3	女装/女士精品 > 毛针织衫	8.74%
4	女装/女士精品 > 毛衣	7.98%
5	女装/女士精品 > 卫衣/绒衫	5.36%
6	女装/女士精品 > 连衣裙	4.59%
7	女装/女士精品 > 牛仔裤	4.47%
8	女装/女士精品 > 半身裙	3.01%
9	女装/女士精品 > 棉衣/棉服	2.39%
10	女装/女士精品 > 休闲裤	2.15%
11	其它	6.70%

竞店1 价格带	支付金额占比
0-70元	0.00%
70-150元	0.05%
150-250元	0.43%
250-400元	12.02%
400-700元	28.98%
700元以上	58.52%

Figure A.3: Traffic Module Interface

Notes: This is a screenshot that illustrates the web interfaces of the paid competition module in the data analytics tool.

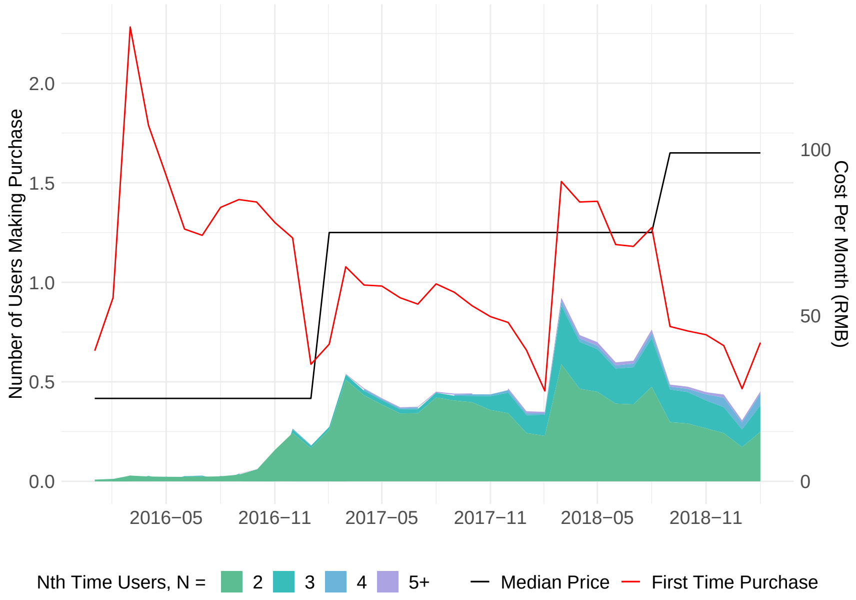


Figure A.4: Adoption and Price Change of Market (Pro) Module

*Notes:*number of orders placed for the Market Pro module. Red line depicts number of orders placed by first time users, while the colored area graph differentiate renewal purchases based on the number of renewal. Number of users are re-scaled.

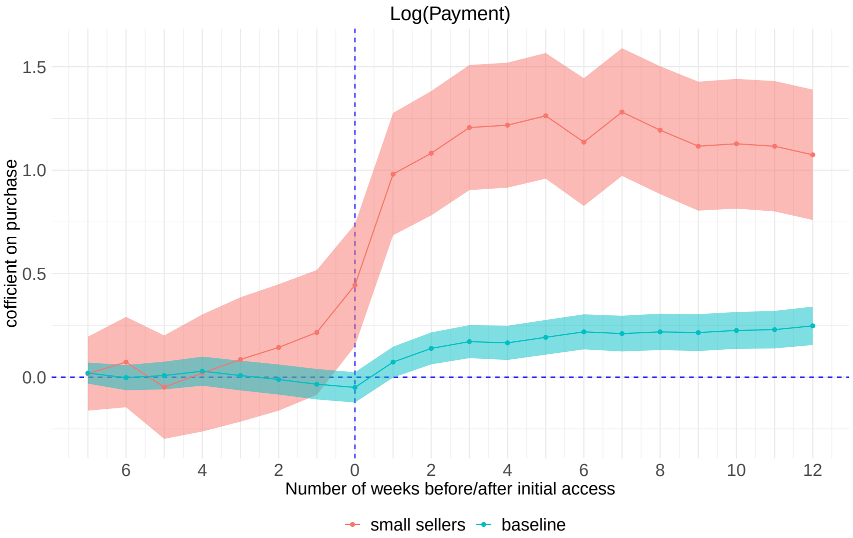


Figure A.5: Effects of Acquiring Own Store Traffic and Customer Data on Log Sales

Notes: the graph shows coefficient γ_k for $k = -11, ..., 11$ where $k = 0$ is the week of adoption, separately for large and small stores regarding adoption of traffic module. A store is small if week sales at week $t - 12$ is in the bottom quarter. Dependent variable is log sales in the previous week. All regressions include controls for stores characteristics, performance and actions profile in the pre-adoption period.

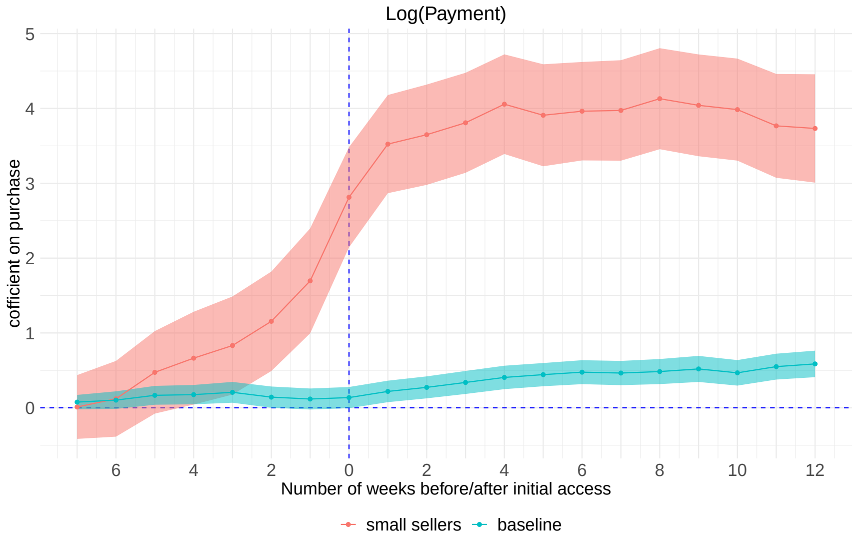


Figure A.6: Effects of Acquiring Top-Ranked stores and Products Data on Log Sales

Notes: the graph shows coefficient γ_k for $k = -11, ..., 11$ where $k = 0$ is the week of adoption, separately for large and small stores regarding adoption of Market (Pro) module. A store is small if week sales at week $t - 12$ is in the bottom quarter. Dependent variable is log sales in the previous week. All regressions include controls for stores characteristics, performance and actions profile in the pre-adoption period.

B Estimation Result

Table B.1: Estimates of Main Parameters

Parameter	Estimates	Std Err
α_0	-9.09e-4	0.17e-4
$\alpha_{\text{(normed) log size}}$	5.609e-4	0.04e-4
$\alpha_{\text{(normed) open time}}$	-1.93e-4	0.01e-4
σ_γ	4.30	0.81
$\sigma_{\text{sales ind}}$	0.27	0.00
$\sigma_{\text{log sales}}$	-0.56	0.00
$\rho_{\gamma,t}$	0.78	0.06
$\theta_{\gamma,k=\text{top rankings,open time}}$	1.25	0.12
$\theta_{\gamma,k=\text{agg demand,open time}}$	1.11	0.03
$\theta_{\gamma,k=\text{competitors,open time}}$	-1.76	0.17
$\xi_{\gamma,k=\text{top rankings}}$	0	-
$\xi_{\gamma,k=\text{agg demand}}$	0.64	0.22
$\xi_{\gamma,k=\text{competitors}}$	1.88	0.15
$\rho_{k\text{:top rankings \& agg demand}}$	0.02	0.05
$\rho_{k\text{:top rankings \& competitors}}$	-0.42	0.16
$\rho_{k\text{:competitors \& agg demand}}$	-0.16	0.07
$\xi_{j=\text{mkt pro}}$	0	-
$\xi_{j=\text{mkt std}}$	-1.97	0.02
$\xi_{j=\text{comp}}$	-1.36	0.00

C Matching Methodologies and Details

The matched samples are constructed using a random sample of about 57,000 stores in industries related to consumer electronics covering the period of April 2017 to December 2018. For each product module, i.e. market (std), market (pro), traffic and competition, matched samples are constructed separately from the same underlying population. For each specific product module, stores are classified into three categories: 1) non-users are stores that have never purchased the module in the sample period; 2) first-time users are stores that have purchased the module for the first time during the sample period and 3) returning users are stores that have purchased the product module in the past that may or may not renewed during the sample period. The analysis on matched samples focus on comparing non-users and first-time users. Therefore, when constructing the matched sample, returning users are excluded from the population. Notice that stores may have different status with respect to the particular product module of interests. That is, some stores might be returning users for traffic module but are first-time users for market (std) module, in which case these stores will be excluded in construction of the traffic matched sample, but will be included for the construction of the market (std) sample.

For each specific product module, the matched sample is constructed in the following steps. First, stores that are first-time users are categorized into different cohorts based on the week t_0 when purchase is made, where $t_0 \in \{2017-12, ..., 2018-40\}$. Exact number of cohorts as well as size of the cohorts vary by module. The most popular module market (std) have 648 qualified first-time users while the least popular module competition only has 158 users. For each cohort, the matched sample is constructed by selecting comparable stores in the non-users sample that are comparable along the following dimensions:

- *basic characteristics*: rating, days opened, type, industry

- *performance*: sales, number of orders, quantity sold, traffics and conversion rates
- *strategies*: frequency of changing product title, adding/removing products

For variables on performance and strategies, all are matched from week $t_0 - 11$ to week t_0. To be specific, the matched sample is constructed by first running the following specification

$$Purchase_i = \alpha + \sum_{t=t_0-11}^{t_0} X_{it}\beta_t + Z_i\gamma + \epsilon_i$$

where X_{it} include variables on performance and strategies in each particular week t as described above and Z_i includes all time-invariant characteristics. $Purchase_i$ is a dummy that equals 1 if the store purchase the module in week t_0 and equals 0 if stores are non-users. Notice that first-time users are not included for construction of the matched sample except for being used as the target of the match for the cohort they belong to. Given the specification, the matched sample is constructed by selecting from the non-users with highest purchase likelihood. Cohort size of the matched sample is the same as cohort size of the first-time users' sample for the specific cohort t_0. The final matched sample for a product module is consists of all cohort-specific matched sample for that product module. Cohort (t_0) dummies are included in the analysis for the matched sample to control for cohort-specific effects. By construction, first-time users and their matched non-users counterpart should behave similarly along all matched dimensions (performance, strategies) up till week t_0 due to the requirements applied in the matching process.