Estimating demand for zero priced products and the value of personal data

Farasat A.S. Bokhari, Michael E. Kummer

School of Economics, University of East Anglia, Norwich NR4 7TJ, UK

Centre for Competition Policy, University of East Anglia, Norwich NR4 7TJ, UK

September 28, 2020

Abstract

In this project we aim at applying structural IO methodology to quantify the value of private data in a demand estimation model. Specifically, we propose to develop an approach that adapts standard demand estimation models to understand how individuals value privacy and information in the context of Apps and how to extend the analysis to merger analysis for zero-priced products. Our focus lies on developing a methodology that can be universally applied to digital markets in which data is being extracted from users. Such a technique could be a valuable asset, because it allows to estimate the value of personal data and to model policy counterfactuals, such as restricting firms access to user data or increasing consumer protection. These are important interventions that are currently being considered by regulators around the globe. We model data alongside price as a second currency in a BLP model with random coefficients for both product dimensions (money and data). As a proof of concept, we apply our methodology to data from Androids mobile app market in 2012. In this market, two models prevailed: a first group of mobile apps was available for installation after consumers agreed to pay the apps price. A second group of apps was offered for free, but these apps typically requested access to more personal data of the user. We leverage the existence of both business models in our identification strategy. Our approach allows to quantify the value of user data and enables the evaluation of counterfactual policy experiments.

Key words:

JEL Classification:

1. Motivation

By the end of 2012, the Google Play Store offered approximately 400,000 apps, and in 2018 the number increased to 2.5 million. This growth illustrates the dynamic character of the industry, which features many apps that grow at different rates, and has an extraordinary amount of entry and exit, that creates considerable variation in the competitive environment across submarkets and over time. The industry consists of 40 subcategories (such as 'Health', 'Weather', 'Action', 'Music', 'Communication' apps) which cover around 5000 submarkets.

Broadly speaking with our paper we hope to contribute the literature which analyzes the value of private data. Within this literature, several theoretical models analyze the role of private information in online markets. In such models, firms use their knowledge about an agent's preferences to price discriminate (Taylor et al., 2010, Wathieu, 2002). Also, firms can use customer information, such as purchase history, to charge personalized prices in electronic retailing settings (Acquisti and Varian, 2005, Conitzer et al., 2012, Taylor, 2004).

Data is clearly valuable to developers, because they can monetize their products via four main channels (for a discussion, see Kummer and Schulte (2016)). According to AppBrain (2016), around 20 percent of the apps in the Google Play Store are paid apps, whereas the remaining 80 percent of apps are free.[1] Alternative revenue channels are in-app advertising, in-app purchases, and data trade. Data sharing for monetization purposes is common in the mobile app industry, as in many online markets (see, e.g., Woodcock, 2017 and references therein). Christl and Spiekermann (2016) and OECD (2013) survey several studies which show this pattern and provide a brief discussion of the data-sharing business model including the three main channels used to generate revenues from data: (1) The most important channel is using information about users to sell targeted advertising. This channel has arguably been the powerhouse of the market for free apps, representing a sizable industry.[2] (2) Another relevant channel of data flows arises when developers trade their data for valuable third-party services.[3] (3) Finally, app developers can trade their data with third parties in several other common ways. For example, they can exchange their data for direct monetary benefit. Such data can be purchased by known industry analysts such as *Alexa.com* or by lesser-known data aggregators.

[1] Android's app developers receive 70 percent of the app price, and 30 percent goes to distribution partners and operating fees (see `https://support.google.com/googleplay/android-developer/answer/112622?hl=en`).

[2] Peer-reviewed estimates are not available. Existing estimates range from several hundred million to several billion US dollars (OECD (2013) based on Beales (2010)). The value in this industry is generated from collecting and aggregating individuals digital traces from various sources (e.g., an app), in order to profile users' location (Dubé et al., 2017, Fong et al., 2015) and behavior for targeted advertising.

[3] An example of such services is app analytics: it helps developers gain insight on who uses their app when and where, along with which other processes on the phone are simultaneously being used. Combined with the developers own knowledge of the users in-app behavior, this can be a useful input for improving the app or, again, for advertising purposes.

However, it is less clear how consumers evaluate the data in this market. Acquisti et al. (2016) provide an excellent overview over this literature, and we will selectively cite directly relevant empirical or theoretical research here. Personal information can be used in direct marketing, which may result in costly efforts to avoid ads (Hann et al., 2008, Johnson, 2013b). Increasing the cost of anonymity can benefit consumers, but only up to a point, after which the effect is reversed (Taylor et al., 2010). A series of empirical studies analyzes how privacy policies affect users of social networks or the success of targeted advertisement (Aziz and Telang, 2015, Goldfarb and Tucker, 2011, Johnson, 2013a, Tucker, 2012, 2014). They show that restricting the use of private data in advertising reduced targeting effectiveness, which resulted in lower revenues for the content site but also highlighted the important effect of privacy policies on consumer behavior. Taken together, these models find that reduced privacy is disadvantageous for consumers.

A few selected papers analyze the role of private information *in app markets.* The existing empirical research on the value of private data in mobile markets is based on experimental and survey data, and most studies focus exclusively on the demand side (see e.g. Savage and Waldman (2014) or Egelman et al. (2013)). An exception are Kummer and Schulte (2019), who exploit observational app data and analyze both the demand and the supply side. Overall, these studies confirm the value of data both for developers as well as the value of privacy for customers. However, none of these papers have provided an approach to structurally estimate the value of data to consumers and to use it in policy counterfactuals. We propose to close this gap.

2. Data and Implementation

In what follows we describe how we want to implement the estimation.

2.1. Implementation. In the standard discrete choice models, an individual derives utility from consuming product j from a choice set J, and where the utility is a function of the price and other characteristics. Thus,

$$\begin{aligned} &U_{njt} = \alpha_n(y_{nt} - p_{jt}) + \tilde{\mathbf{x}}_{jt}\boldsymbol{\beta}_n + \xi_{jt} + \epsilon_{njt}, \text{ where} \\ &n = 1, \ldots, N, \qquad j = 0, 1, \ldots, J, \qquad t = 1, 2, \ldots, T. \end{aligned} \tag{1}$$

In the equation above, U_{njt} is the utility of person n living in market t and choice product j and has an income of y_{nt}. Similarly, p_{jt} is the price paid for product j, $\tilde{\mathbf{x}}_{jt}$ is the set of k characteristics of the selected product (i.e, $\tilde{\mathbf{x}}_{jt}$ is a k vector). In the set-up above, the coefficients α_n and $\boldsymbol{\beta}_n$ are the marginal utilities of income and of the various characteristics. The terms ξ_{jt} are the unobserved (to the econometrician) product characteristics, and ϵ_{njt} is the idiosyncratic error term, typically assumed to be distributed as iid extreme value. Setting

$$\boldsymbol{\beta}_n = \boldsymbol{\beta} \qquad \text{and} \quad \alpha_n = \alpha \qquad \text{for all } N \tag{2}$$

results in the simple logit demand model, while if we do not make such a assumption, we obtain the random coefficients demand model, and estimate it following the methods set out in Berry et al. (1995) (henceforth just BLP).

For the case of Apps, we can partition the product characteristics into those that extract information from the consumers, for the purpose of monetizing this information, and which presumably lowers the utility of a given consumer, and other characteristics of the product that increase the utility. Thus if the information extraction of an App is captured as a scalar value, then we can write

$$\tilde{\mathbf{x}}_{jt} = [i \qquad \mathbf{x}]_{jt},$$

where i is the scalar variable for information extraction, and associated with a negative marginal utility b_1, and $\mathbf{x}$ is a vector of all the remaining product characteristics with marginal utilities given by $\boldsymbol{\beta}_2$. For the simplifying case of Equation 2 (and also ignoring the t subscript), we can then write the utility as just

$$U_{nj} = \alpha(y_n - p_j) + i_j\beta_1 + \mathbf{x}_j\boldsymbol{\beta}_2 + \xi_j + \epsilon_{nj}, \tag{3}$$

which under the usual distributional assumptions gives the plain vanilla logit demand model

$$\ln(s_j) - \ln(s_0) = \alpha(p_{jt}) + \beta_1 i_j + \mathbf{x}_j\boldsymbol{\beta}_2 + \xi_j \tag{4}$$

and where s_j and s_0 are the quantity shares of product j and the outside option.

A significant part of the research work would be spent in understanding the market to construct the outside option and obtaining valid and relevant instruments for p and i as they are likely to be correlated with the unobserved product characteristics ξ_j. Assuming these practical problems can be overcome, the model can be estimated using a simple logit specification, and where the coefficients β_1/α provides a money metric for (dis) utility of information extracted by an App. The model can be enhanced by allowing consumers to differ in their marginal utility of income and of the informational characteristics, and accordingly estimate a random coefficients model (i.e. do not impose restrictions in equation 2 above).

In our application, we will also model supply side, where an App developer (who may have multiple Apps) takes the decision to set prices and level of information extracted via any given App. For instance, for the f firm with a total of L products, the firms profit function would be

$$\max_{p_l, i_l} \Pi_f = \sum_{l}^{L} [(p_l + \gamma_l i_l) - c_l] q_l \tag{5}$$

and its first order conditions would be in the two strategic variables p and i (i.e., $\partial\Pi_f/\partial p_l = 0$ and $\partial\Pi_f/\partial i_l = 0$ for all l), and where c_l is the marginal cost set to zero and γ is the monetization value of information collected by the firm from all consumers. We would then solve for both the unconstrained Nash equilibrium in prices and information, and compare the solution to constrained problems where some firms set in equilibrium $p_j^* = 0$. The constrained would be based either on p_j^* crossing a threshold, else it is not worthwhile collecting from consumer (for instance if there is a price collection fee that the firm must pay), or by introducing an additional product characteristic which shows ads to the consumers in lieu of charging a price.

The merger simulations would be based on estimated model parameters and backed out values of γ_l based on competition mode (Nash in prices and information). With the parameters in hand, we can then modify the ownership structure to simulate new prices and information values post a hypothetical or a real merger. Equally important, we can also then simulate changes in firm profits and in consumer welfare even when two firms selling zero-priced products merge.

2.2. **Instruments.** As modeled in equation 5, the two main variables that a firm uses to maximize profits are price and information and hence their optimal values determined in a Nash equilibrium competition model would be functions of demand. Accordingly, we treat them as endogenous variables and look for instruments that would be correlated with their values but not necessarily with demand.

The first is the size the app. Size in bytes would be a function of the lines of code and would be correlated with programmer time and/or complexity of the app, and hence would be a development cost.[4] An end user would typically not care about size as most email clients are designed so as not to take up too much space on a device. A second potential instrument is a developer's propensity to monetize information as a business model. Thus, for each reference app j, we would compute the mean value of information index i for other apps by the same developer in the non-email group of apps. The information index in other apps should not affect demand for the reference app but would be correlated with reference apps index value due to monetization strategies that may common across product groups.

2.3. **Data.** The data we will use in order to analyze in our proof of concept is based on the data that are published in Kummer and Schulte (2019). These data contain information collected from one of the two largest platforms for mobile applications. These data were collected since the beginning of 2012 and they hold a rich set of characteristics of most apps available in this market. The information is available at a monthly frequency and includes most of the information which is publicly available for each app. Thus, we have information on current prices, the category an app belongs to (e.g. 'games', 'news & magazines', 'communication', 'books', etc.), and the number of and the average of ratings of an app. Additionally we observe several indicators for the way an app is presented in the app store ('length of description', 'number of screenshots', 'video available', etc.). Furthermore, we

[4]We note that code length is not a marginal cost which is a concern when using it as instrument in the price equation. Ongoing efforts consist of identifying other suitable measures for marginal cost or developing an adaptation of the approach that is based on average cost considerations.

have data on the total number of installations and sales for each app. This information exists in a discrete form (17 levels, e.g. 1-5 installations, 6-10 installations, 11-50 installations, etc.). Another group of information available are characteristics of an apps close competitors. Collecting such information is possible since for each application information is given on related applications via a section 'users who viewed this also viewed'.

Beyond the variables that are related to sales, the information on the precise number and type of permissions an app uses is crucial to our set of research questions. We observe 136 different rights, which an app can request (e.g. "Access to the precise location", "Downloading data without permission", etc.). If an application uses such permissions, this is indicated during the installation process within the application. Moreover it is permanently displayed on the applications page on the platform where it can be downloaded.

The raw data set is based on the information collected since April 2012 and contains weekly information on nearly all available applications (N= 300,000 and T = 31 weeks). A second data set, which features daily information on apps which appeared in the "Top-512-Newcomer" list during the 19th of October 2012 until the 2nd of December 2012 can be used as well.

3. Implementation

We now turn to the details of how we implement the estimation procedure. We first discuss how we arrive at the focal category of email-clients. Next, we discuss how we implement the outside option, the demand measure and how we compute an app's market share. We then explain the adopted estimation of the model, provide basic descriptives of the focal category and discuss the computation of the instrumental variables.

Category selection: As proof of concept, we develop our demand model using a group of related apps as defined by Google's categories or subcategories. Our criteria for selection are sufficient cross-sectional variation (i.e., across individual products) in (i) price p_j (ii) information i_j and (iii) ability to identify some exogenous source of variation in these variables in order to be able to estimate model parameters. Additionally, we focus on applications where an outside option is easy to identify in order to be able to compute the share of the outside option (s_0) which varies over markets. Our initial investigation showed that e-mail client apps may meet these criteria. Hence we chose e-mail clients for our first implementation, but we emphasize, that our approach is designed to work with any other suitable group of apps.

Outside Option for email clients: Android devices require users to create a Google account and the device comes equipped either with a G-mail client already installed, or provide access to the email servers via integrated browsers. However, users may want to download and use other email clients with different attributes that allow them to communicate with other commercial or work

Table 1. Main Variables in the email-cluster

	mean	sd	p5	p25	p75	p95
Δ Ratings	9.98	32.52	0.00	0.00	4.00	50.00
Δ Installations	4633.56	28733.86	0.00	0.00	0.00	4500.00
Ratings	271.21	1227.46	1.00	4.00	53.00	901.00
Installations	32344.28	1.2e+05	30.00	300.00	7500.00	75000.00
#_TotalPerm	6.27	5.19	1.00	3.00	8.00	19.00
Price — *Price* > 0	1.75	1.42	0.68	0.75	2.17	4.55
D_Privacy	0.78	0.41	0.00	1.00	1.00	1.00
#_Privacy	2.02	1.87	0.00	1.00	3.00	6.00
#_CleanPerm	4.25	3.75	0.00	2.00	5.00	13.00
D_PrivCatSpec	0.04	0.20	0.00	0.00	0.00	0.00
*D_MTurkEP*2	0.64	0.48	0.00	0.00	1.00	1.00
D_Google	0.67	0.47	0.00	0.00	1.00	1.00
D_Sarmaetal	0.76	0.43	0.00	1.00	1.00	1.00
D_ID	0.33	0.47	0.00	0.00	1.00	1.00
D_Location	0.10	0.30	0.00	0.00	0.00	1.00
D_Communication	0.36	0.48	0.00	0.00	1.00	1.00
D_Profile	0.63	0.48	0.00	0.00	1.00	1.00
D_Internet	0.86	0.35	0.00	1.00	1.00	1.00
D_Ads	0.54	0.50	0.00	0.00	1.00	1.00
Price	0.72	1.25	0.00	0.00	0.99	3.72
Average Rating	3.74	0.85	2.30	3.30	4.30	5.00
Size	845.93	1202.43	46.00	135.00	901.00	2900.00
Length Description	927.94	801.50	185.00	335.00	1291.00	2915.00
Number Screenshots	3.43	2.05	0.00	2.00	4.00	8.00
Dummy: Video	0.11	0.32	0.00	0.00	0.00	1.00
Dummy: Top-Dev.	0.00	0.00	0.00	0.00	0.00	0.00
Apps by Developer	12.42	32.46	1.00	2.00	12.00	35.00
Dev. Avg Installations	75811.41	1.5e+05	75.00	1102.50	75000.00	3.0e+05
Observations	125					

related email servers (e.g. Yahoo, Hotmail, Exchange etc.) or provide alternative integration with other utility software such as calendar or direct messaging. Thus, a ready made outside option is the G-mail client app and/or browser, and the size of the market in any period is the total number Android devices (phones and tablets) which can be estimated from the total number of installed or downloads for G-mail clients or other outside data on sales of devices. Changes in this number from period to period would then give us a measure of new android devices per period Q_t, which would be the total market for that period. Downloads or ranks q_{jt}, our two potential measures of quantity (discussed further below), for email app j within the month t then gives a measure of the share as $s_{jt} = q_{jt}/Q_t$ from which we can construct outside share as $s_{0t} = 1 - \sum_j^J s_{jt}$.

Using Ratings as Demand Measure: We have two demand measures at our disposal - a categorial measure of installations and ratings. While both of these measures are highly correlated our measurement of demand is imperfect for both. The categorial measure of installations reflects actual downloads of the app, but the measure is not very precise. Typical categories are 50,000-100,000 or 100,000-500,000 installations, but within, say, the latter category any value is possible. Hence, although using installations is generally preferable, the lack of precision in the measurement implies that apps stay in the same band for several periods before jumping to the next band. This means that any given app j has no visible variation in most periods of time. This problem does not exist for ratings, which are measured with greater precision and hence more variation over time! Hence, while ratings are not exactly the variable we are actually interested in - Installations - they are a highly correlated alternative that is available with greater precision. Hence, we use ratings as our primary demand measure, and we rescale the value for the outside option accordingly. For email applications that is not a problem, because we can directly use the ratings of Gmail. Alternatively we can rescale installations of Gmail by the installations/ratings ratio that is prevalent for email clients.

Adjusting the model for Gmail as outside option: Because we consider Gmail as the outside option, the value of the outside option is not 0 in our context. As a result, the standard logit model needs to be adjusted. As stated above, the standard version of the model would estimate the following model:

$$\ln(s_j) - \ln(s_0) = \alpha(p_{jt}) + \beta_1 i_j + \mathbf{x}_j \boldsymbol{\beta}_2 + \xi_j$$

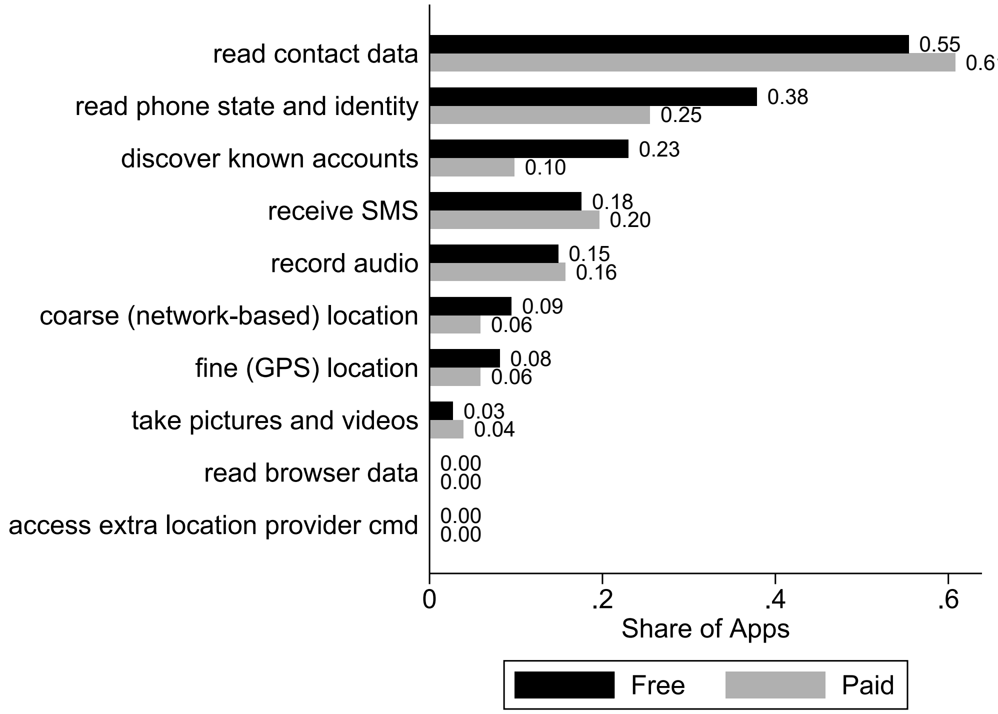
where s_j and s_0 are the quantity shares of product j and the outside option. However, to account for the non-zero outside option, and we estimate the model in terms of how an app differs from the outside option. For email clients gmail is the outside option on Android, and hence, after accounting for the non-zero outside option we estimate:

$$\ln(s_{jt}) - \ln(s_{0t}) = (\mathbf{x_{jt}} - \mathbf{x_{0t}})\beta - \alpha(p_{jt} - p_{0t}) + (\xi_{jt} - \xi_{0t}). \tag{6}$$

Baseline Descriptives: With this plan in the background, the subsequent tables and figures provide information on the data that is available on email clients. Table 1 shows the distribution of the main variables in the selected category of email clients. A distinguishing feature of email clients is the large share of apps with a positive price (51 apps or over 40%). Figure 1 provides a comparison of how the most commonly used intrusive permissions are distributed between free and paid apps. Table 2 compares the main variables in the dataset for free and paid email clients. All these comparisons are currently available on the cross section, but ongoing work is putting together the corresponding panel datset.

Instrumental Variables: We model price and privacy as profit-maximizing choice. In other words, because we allow developers to optimally choose their pricing and privacy strategy in equilibrium, they are determined inside the model. While there are several endogeneity concerns at work, the most concerning one in this industry would be that developers cannot easily anticipate how successful their app will be, and hence might receive a positive or negative demand shock, once they published their app. A widely observed behavior by developers is to launch their app, and add a monetization strategy if they realize that the app is successful. Such a strategy could be based on raising the price, or on adopting a data-driven strategy of selling ads. In such a scenario,

Figure 1. Prevalence of permissions in the category.



Notes: Prevalence of permissions in the category of mail clients. Based on Kummer and Schulte (2019).

price and/or data-collection (i.e. privacy) depend on an app's success and become endogenous. A simple OLS regression of market share on price/privacy that takes the determinants as exogeneously given would be biased. Instead, we apply an instrumental variable approach that relies on finding exogenous instrumental variables (IVs) that are predictive of the strategic (and hence endogenous) variables. In principle, we can use three different types of IVs:

(1) Our first strategy relies on finding cost shifters, that are sunk by the time the developer learns how successful their app will be. We use code size as a direct predictor of price.
(2) A second strategy could exploit the fact that competitors' characteristics are observed for four competitors. We could then generate set of instrumental variables z_comp that uses the characteristics of an app j's competitors to predict the characteristics of j. However, while the competitor's behavior is highly predictive, it remains doubtful whether competitor behavior is really exogeneous, and whether we can rule out that j's success influences its competitor's choices, and common underlying drivers that drive the choices of both app j and its competitors. Moreover, it is plausible that developers consciously adjust to their competitors. Hence, we deem competitor characteristics not ideal and refrain from using them.

Table 2. Comparison of Free and Paid Variables

	Cross Section			
	0		1	
	mean	sd	mean	sd
Δ Ratings	13.30	39.01	5.16	19.00
Δ Installations	7820.27	37110.84	9.71	63.20
Ratings	397.18	1567.10	88.43	305.80
Installations	52721.35	1.5e+05	2777.56	7034.81
$\#_{TotalPerm}$	6.85	6.07	5.43	3.44
$D_{Privacy}$	0.76	0.43	0.82	0.39
$\#_{Privacy}$	2.20	2.16	1.76	1.32
$\#_{CleanPerm}$	4.65	4.24	3.67	2.83
$D_{PrivCatSpec}$	0.04	0.20	0.04	0.20
$D_{MTurkEP2}$	0.61	0.49	0.69	0.47
D_{Google}	0.65	0.48	0.71	0.46
$D_{Sarmaetal}$	0.73	0.45	0.80	0.40
D_{ID}	0.38	0.49	0.25	0.44
$D_{Location}$	0.12	0.33	0.06	0.24
$D_{Communication}$	0.32	0.47	0.41	0.50
$D_{Profile}$	0.62	0.49	0.65	0.48
$D_{Internet}$	0.88	0.33	0.82	0.39
D_{Ads}	0.65	0.48	0.39	0.49
Price	0.00	0.00	1.75	1.42
Average Rating	3.60	0.83	3.94	0.84
Size	753.39	971.68	980.20	1475.17
Length Description	917.38	807.70	943.25	800.17
Number Screenshots	3.42	1.93	3.45	2.22
Dummy: Video	0.14	0.34	0.08	0.27
Dummy: Top-Dev.	0.00	0.00	0.00	0.00
Apps by Developer	13.41	41.22	11.00	11.39
Dev. Avg Installations	96355.44	1.4e+05	46002.44	1.5e+05
Observations	125			

(3) The third approach uses a developers' choices on their other apps (in other categories), to generate instrumental variables of the group *devcat*. The idea of this approach is that the developer's previous choices and experience from earlier apps are predictive of the design and monetization strategy that a developer will use on the app in question, without being a function of the app in question. Especially correlations with apps in other product categories are more likely to be a result of the developer's underlying skill set and programming approach, rather than some ongoing market dynamics. The most relevant two examples of this

IV-set are the permissions a developer requires in other apps, or the prices the developers charge for their other app.

4. Results

We now turn to the results for email-clients in Table 3. The table shows five columns. The first three columns estimate the model of interest, while columns 4-5 show the first stage regression that is associated with the instrumental variable estimation. In the first column we show the results of the OLS estimation, which is expected to lead to biased coefficients. In columns 2 and 3 we show two variants of the 2SLS IV-estimation. The instruments in column 2 are a developer's pricing and privacy strategies in their apps from other categories. In column 3 we also use a developer's privacy strategy from other apps, but use the cost-shifter instrument (code-length) from apps in other categories to predict price. The first-stage regressions in column 4 and 5 reveal that the instruments are relevant to predict privacy, but only the cost-shifter instrument is relevant for predicting price. Hence, the first-stage regressions suggest that the IV specification in column 3 is preferable to the specification in column 2.

When looking at the results we first look at the benchmark OLS specification, which, ultimately, cannot be trusted. Indeed, the OLS estimation suggests that users prefer intrusive apps, which corroborates the concern that developers might up their permission requests after their app turns out to be successful. The results in IV(2) uses avg. price and avg privacy, and the coefficients become very large and statistically insignificant, which is suggestive of a weak IV problem. Indeed, the first stage regression in column 4 confirms, that the instrument that uses a developer's average price is not relevant. IV(3) uses avg. privacy and avg. *code size* as a cost-shifter instrument. The first stage regression in column 4 highlights that the size based IV is more relevant, and the coefficients, indeed appear to be more plausible than in the IV2 or the OLS-estimation.

Altogether, a word of caution is in place, before further using these results to compute substitution elasticities or the welfare implications of mergers. First, the results on email clients are not very robust and exhibit a large amount of variability, at least at the current stage. We noted a potential issue of multicollinearity in the instruments, which might explain why, even in the preferred specification, we find relatively large and and imprecisely estimated coefficients. A second major source of concern is the standard logit approach is the massive loss of observations that we are faced with. Starting from 651 observations we loose 130 when computing installation/ratings growth, that is estimated as the first difference in the stock of ratings. Next, an additional 220 observations are characterized by 0 market share, because they did not get any ratings in period t. Hence, these observations are also not used in the logit, and ultimately we have to estimate the model with less than 50% of the observations in the data. In what follows we will explore improvements to the current approach that aim at providing a remedy to these issues.

Table 3. Results based on the email-client category

	OLS (1)	IV (2)	IV (3)	First Stage (4) Price (x2)	Privacy (x3)
x_2 - price	-0.211*	-6.582	-0.980*		
	(0.0648)	(22.80)	(0.582)		
x_3 - privacy	0.165**	-2.097	-0.465		
	(0.0790)	(7.084)	(0.373)		
x_5 - # clean permissions	0.0625	0.752	0.232*	-0.00809	0.298*
	(0.0397)	(2.283)	(0.133)	(0.0281)	(0.0224)
x_6 - average ratings	0.890*	2.936	1.024***	0.272**	0.136
	(0.182)	(7.849)	(0.298)	(0.117)	(0.133)
x_8 - length description	-0.000133	0.00286	0.000381	0.000367*	-0.0000479
	(0.000149)	(0.0101)	(0.000308)	(0.000213)	(0.000117)
x_9 # of screen shots	0.203*	-0.109	0.0900	-0.0290	-0.0191
	(0.0721)	(0.892)	(0.0801)	(0.0499)	(0.0479)
x_{10} # of apps by developer	0.00667	-0.0131	-0.00271	0.000195	-0.00464*
	(0.00605)	(0.0502)	(0.00577)	(0.00163)	(0.00162)
x_{11} 0/1 - Video	0.273	-0.490	-0.0315	0.127	-0.743*
	(0.319)	(2.796)	(0.438)	(0.339)	(0.224)
zx_2: avg. price (Dev) IV for x2				0.0380	-0.109
				(0.170)	(0.0703)
zx_3: avg. privacy (Dev) IV for x3				-0.0959	0.378*
				(0.0614)	(0.0686)
zx_7: size IV for x2				0.000178**	-0.0000619
				(0.0000877)	(0.0000686)
Constant	-2.772*	-13.17	-10.07*	0.644*	-3.918*
	(0.0292)	(14.39)	(1.209)	(0.387)	(0.363)
Week Dummies	YES	YES	YES	YES	YES
N	283	278	278	284	284
R^2	0.289	.	.	0.118	0.585

Standard errors in parentheses;* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

5. Alternative Approaches

In the previous section we implemented a structural logit estimation that focused on estimating parameters for both app price and privacy settings in the domain of email clients, and we showed that the approach suffers from serious limitations. Most specifically the standard methodology induces a serious loss of observations, given the “long tail” with of apps with zero market share. As a result we faced challenges to find sufficiently predictive instruments, and are faced with relatively imprecise coefficient estimates. To solve these issues we are proposing to explore two improvements! First, we attempted a broader analysis across categories, that considers each category a market of it’s own, and thus increases statistical power. Second, in ongoing work, we would like to explore a modified approach that allows us to leverage the information contained in an app with 0 market share.

5.1. **Estimation across all app categories.** In this section we discuss how to adopt the approach that we sketched above to an estimation across all app-categories that considers each app category as an independent market. Table 4 in the Appendix shows the distribution of the variables in all app categories. Most notably, we can use many more apps in this estimation, and rather than having only 5 periods as 5 markets, we can use thirty categories in 5 periods to obtain 150 markets. However, several adjustments need to be made to estimate the model across multiple categories.

First, Gmail can no longer be the universal outside good, and hence we need a new definition of the outside good. To account for this issue, we define the outside good to be "not using any app from category j" in our analysis across categories. This assumption seems plausible if we are willing to assume that nobody is compelled to use an app from each category. An alternative estimation strategy could seek to identify universal benchmark apps for each category. [5] Given the absence of a reference app, we define the Market Size for each category based on two ingredients: the number of Android users and the number of total installations in the category. We then define market size as the sum of new Android phones and existing Android Users that did not use apps from a given category until period $t-1$.[6]

The results of this estimation approach are shown in Table 5 in the Appendix. As before Column 1 estimates the benchmark OLS model, which is expected to give biased coefficients. Columns 2 and 3 show the results of 2 alternative IV-estimations (with and wo category dummies) and columns 4 and 5 show the corresponding first-stage regressions. As before, we are instrumenting price with the developer's average price on their other apps. Whenever it is possible we use only apps from other categories, but if a developer has only apps in the same category, we use those to compute the instrumental variable.[7] Privacy is, analogously, instrumented with a developers' average permission requests on their other apps from other categories. These instrumental variables are highly predicitive of the endogeneous variables, as can be seen in columns 4 and 5.

Unlike in the previous table, columns 2 and 3 don't differ in their use of instruments, but in whether they use of wave dummies and category dummies. In column 2 we do not use such dummies and in column 3 we do. Between these two columns the privacy coefficient is estimated to lie between -0.13 and -.066, and it diminishes by 50% when controlling for category and wave (in column 3). The estimated parameter for price analogously diminshes by 50%, when introducing the category dummies in column 3. In general the coefficients seem to lie within more reasonable bounds in the specification that leverages the data from all categories.

[5] For example, facebook might be considered the benchmark app for social networks.

[6] Existing Android users are computed based on all Users in previous month - Category Adopters in previous month (total category installations.

[7] Apps of single-app developers have been randomly combined into pseudo-app developers with apps of similar properties in adjacent categories.

5.2. **Leveraging the information contained in 0-demand apps.** The second limitation of the estimation above is that the majority of observations is discarded due to 0 demand. Our idea for improving the estimation strategy is based on the observation a 0 market share is informative of very low demand. If users do not install an app given its quality, price and privacy settings we can infer that the bundle is not atractive to current consumers. However, a standard logit will drop the observations without considering them. We argue that it is likely to be informative to consider which app-characteristics fail to attract any demand and suggest to find ways for including this information in the estimation.

Ongoing work is exploring how this insight might be leveraged systematically, but sevearl strategies come to mind. First, we can add an analysis of the sample selection. However, that does not allow to include the zero-demand apps in the estimation. An alternative would be to account for the fact that every rating typically represents several installations and attributing a value larger than 0 to zero installations. A third approach might seek to combine the demand estimation with a Tobit approach that allows to include the censored observations.

6. Conclusions

We have implemented a first conditional logit estimator, based on email-clients. In this approach we use the industry standard (gmail) as non-zero outside option. We implement several IV approaches and apply them in estimations, and thus provide a useful first step towards implementing digital merger analysis. The approach is not free of limitations, however. Most specifically, when working only with a relatively small cluster of apps, it is relatively challengeing to find predictive instruments. Second, because app markets are characterized by a long tail with zero market share, the standard methodology induces a serious loss of observations which is tantamount to neglecting valuable data about unsuccessful offers.

Ongoing work is exploring two approaches to improve the results in the main analysis. First, we propose using a broader analysis across categories to achieve a more robust inference (cf. above). Second, we propose exploring how to leverage the information that lies in observing apps with a 0-market share. Further ongoing work can use the estimated parameter simulate merger or other policy changes etc.

References

Acquisti, A., Taylor, C., and Wagman, L. (2016). The economics of privacy. *Journal of Economic Literature*, 54(2):442–92.

Acquisti, A. and Varian, H. R. (2005). Conditioning prices on purchase history. *Marketing Science*, 24(3):367–381.

AppBrain (2016). Google play stats. Available at http://www.appbrain.com/stats/free-and-paid-android-applications.

Aziz, A. and Telang, R. (2015). What is a cookie worth? *Heinz College Working Paper*.

Beales, H. (2010). The value of behavioral targeting. *Network Advertising Initiative*, 1.

Berry, S., Levinsohn, J., and Pakes, A. (1995). Automobile prices in market equilibrium. *Econometrica*, 63(4):841–890.

Christl, W. and Spiekermann, S. (2016). *Networks of Control - A Report on Corporate Surveillance, Digital Tracking, Big Data & Privacy*. Facultas, Vienna.

Conitzer, V., Taylor, C. R., and Wagman, L. (2012). Hide and seek: Costly consumer privacy in a market with repeat purchases. *Marketing Science*, 31(2):277–292.

Dubé, J.-P., Fang, Z., Fong, N., and Luo, X. (2017). Competitive price targeting with smartphone coupons. *Marketing Science*, 36(6):944–975.

Egelman, S., Felt, A. P., and Wagner, D. (2013). Choice architecture and smartphone privacy: There is a price for that. In *The Economics of Information Security and Privacy*, pages 211–236. Springer.

Fong, N. M., Fang, Z., and Luo, X. (2015). Geo-conquesting: Competitive locational targeting of mobile promotions. *Journal of Marketing Research*, 52(5):726–735.

Goldfarb, A. and Tucker, C. (2011). Privacy regulation and online advertising. *Management Science*, 57(1):57–71.

Hann, I.-H., Hui, K.-L., Lee, S.-Y. T., and Png, I. P. (2008). Consumer privacy and marketing avoidance: A static model. *Management Science*, 54(6):1094–1103.

Johnson, G. A. (2013a). The impact of privacy policy on the auction market for online display advertising. Technical report.

Johnson, J. P. (2013b). Targeted advertising and advertising avoidance. *The Rand Journal of Economics*, 44(1):128–144.

Kummer, M. and Schulte, P. (2019). When private information settles the bill: Money and privacy in google's market for smartphone applications. *Management Science*.

Kummer, M. E. and Schulte, P. (2016). When private information settles the bill: Money and privacy in google's market for smartphone applications. ZEW Discussion Papers 16-031, ZEW / Center for European Economic Research.

OECD (2013). Exploring the economics of personal data: A survey of methodologies for measuring monetary value. *OECD Digital Economy Papers*, (220).

Savage, S. J. and Waldman, D. M. (2014). The value of online privacy: Evidence from smartphone applications. Technical report.

Taylor, C. R. (2004). Consumer privacy and the market for customer information. *The Rand Journal of Economics*, 35(4):631–650.

Taylor, C. R., Conitzer, V., and Wagman, L. (2010). Online privacy and price discrimination. Technical report.

Tucker, C. (2012). The economics of advertising and privacy. *International Journal of Industrial Organization*, 30(3):326–329.

Tucker, C. (2014). Social networks, personalized advertising and privacy controls. *Journal of Marketing Research*, 51(5):546–562.

Wathieu, L. (2002). Privacy, exposure and price discrimination. Technical report.

Table 4. Main Variables in all app categories.

	mean	sd	p5	p25	p75	p95
Δ Ratings	21.14	211.87	0.00	0.00	3.00	40.00
Δ Installations	7002.57	2.9e+05	0.00	0.00	0.00	0.00
Ratings	290.08	3555.25	1.00	2.00	37.00	530.00
Installations	62583.00	6.4e+05	30.00	300.00	7500.00	3.0e+05
$\#\_TotalPerm$	4.11	4.14	0.00	1.00	6.00	11.00
Price — $Price > 0$	1.86	3.07	0.69	0.75	2.00	4.57
$D\_Privacy$	0.48	0.50	0.00	0.00	1.00	1.00
$\#\_Privacy$	1.14	1.71	0.00	0.00	2.00	4.00
$\#\_CleanPerm$	2.97	2.82	0.00	1.00	4.00	8.00
$D\_PrivCatSpec$	0.22	0.41	0.00	0.00	0.00	1.00
$D\_MTurkEP2$	0.12	0.33	0.00	0.00	0.00	1.00
D_Google	0.34	0.47	0.00	0.00	1.00	1.00
$D\_Sarmaetal$	0.46	0.50	0.00	0.00	1.00	1.00
D_ID	0.32	0.47	0.00	0.00	1.00	1.00
$D\_Location$	0.27	0.44	0.00	0.00	1.00	1.00
$D\_Communication$	0.08	0.26	0.00	0.00	0.00	1.00
$D\_Profile$	0.19	0.39	0.00	0.00	0.00	1.00
full Internet access $\times D\_Price$	0.12	0.33	0.00	0.00	0.00	1.00
viewnetworkstate_int	0.50	0.50	0.00	0.00	1.00	1.00
Price	0.47	1.74	0.00	0.00	0.50	2.25
Average Rating	3.91	0.97	2.00	3.40	4.70	5.00
Size	2740.13	5751.42	41.00	235.00	3000.00	11000.00
Length Description	745.62	768.49	96.00	283.00	955.00	2435.00
Number Screenshots	3.45	1.91	2.00	2.00	5.00	8.00
Dummy: Video	0.09	0.29	0.00	0.00	0.00	1.00
Dummy: Top-Dev.	0.01	0.08	0.00	0.00	0.00	0.00
Apps by Developer	102.84	350.01	1.00	2.00	33.00	302.00
Dev. Avg Installations	76927.35	6.4e+05	75.00	1002.00	52500.00	2.0e+05
Observations	19195	x 5 periods				

Estimation across all app categories (initial)

This section provides an overview over descriptive statistics and the result when we run the regression of shares (logit) on prices, privacy and other controls

Table 5. OLS and IV Regression of Simple Logit (All Apps)

	OLS (1)	IV (2)	IV (3)	First Stage (3) Price (x2)	First Stage (3) Privacy (x3)
x_2 - price	-0.12^a	-0.33^a	-0.16^a		
	(0.0068)	(0.049)	(0.047)		
x_3 - privacy	-0.042^a	-0.13^a	-0.066^a		
	(0.0081)	(0.020)	(0.018)		
x_5 - # of clean permissions	0.13^a	0.16^a	0.12^a	0.0011	0.26^a
	(0.0045)	(0.0075)	(0.0068)	(0.0032)	(0.0024)
x_6	0.29^a	0.29^a	0.32^a	0.019	-0.099^a
	(0.014)	(0.014)	(0.013)	(0.012)	(0.0088)
Wave Dummies	No	No	Yes	Yes	Yes
Category Dummies	No	No	Yes	Yes	Yes
Observations	32,551	32,354	32,354	32,354	32,354
R-squared	0.120	0.090	0.253	0.065	0.632

Notes. This table is an extended version of regression coefficients shown in the main text. Superscripts a, b, c indicate significance at 1%, 5% and 10%, respectively.

Table 5. OLS and IV Regression of Simple Logit (All Apps)

	OLS (1)	IV (2)	IV (3)	First Stage (3) Price (x2)	First Stage (3) Privacy (x3)
$x_8 (\times 10^{-3})$ - length description	0.17^a (0.012)	0.22^a (0.016)	0.23^a (0.014)	0.17^a (0.0097)	-0.024^a (0.0073)
x_9 # of screen shots	0.092^a (0.0051)	0.097^a (0.0055)	0.096^a (0.0051)	0.038^a (0.0042)	-0.0040 (0.0032)
$x_{10} (\times 10^{-3})$ # of apps by developer	-0.82^a (0.047)	-0.88^a (0.049)	-1.34^a (0.050)	-0.038 (0.043)	-0.53^a (0.033)
x_{11}	0.40^a (0.028)	0.43^a (0.031)	0.18^a (0.029)	0.21^a (0.024)	0.044^b (0.018)
x_{12} - age rating (1-5) (1/0 dummy, 1 if $x_{12} = 2$)	-0.24^a (0.064)	-0.23^a (0.066)	-0.079 (0.061)	-0.068 (0.053)	0.40^a (0.040)
x_{12} - age rating (1-5) (1/0 dummy, 1 if $x_{12} = 3$)	-0.11^a (0.025)	-0.030 (0.035)	0.091^a (0.033)	-0.16^a (0.019)	1.12^a (0.015)
x_{12} - age rating (1-5) (1/0 dummy, 1 if $x_{12} = 4$)	0.16^a (0.040)	0.21^a (0.043)	0.30^a (0.040)	-0.079^b (0.033)	0.60^a (0.025)
x_{12} - age rating (1-5) (1/0 dummy, 1 if $x_{12} = 5$)	-0.22^a (0.047)	-0.20^a (0.048)	-0.081^c (0.044)	-0.016 (0.039)	0.22^a (0.029)
zx_3 - Instrument for x_3				-0.025^a (0.0061)	0.40^a (0.0047)
zx_2 - Instrument for x_2				0.16^a (0.0066)	-0.021^a (0.0050)
Constant	-13.8^a (0.057)	-13.8^a (0.059)	-12.4^a (0.063)	-0.16^a (0.055)	-0.14^a (0.042)
Wave dummies (1,6,11,16,21) (1/0, 1 if week =11)			-0.057^b (0.023)	-0.026 (0.021)	0.0046 (0.016)
Wave dummy (1/0, 1 if week =16)			-0.20^a (0.024)	-0.0028 (0.021)	-0.0057 (0.016)
Wave dummy (1/0, 1 if week =2)			-0.32^a (0.024)	0.0037 (0.021)	-0.020 (0.016)
2.cat-enc			-2.72^a (0.065)	0.45^a (0.053)	0.023 (0.040)
3.cat-enc			-1.89^a (0.049)	-0.041 (0.044)	0.038 (0.033)
4.cat-enc			-2.18^a (0.072)	0.29^a (0.061)	0.29^a (0.046)
5.cat-enc			-1.51^a (0.081)	0.034 (0.072)	0.069 (0.054)
6.cat-enc			-1.76^a (0.054)	0.0089 (0.048)	0.024 (0.036)
7.cat-enc			-1.53^a (0.14)	-0.035 (0.13)	-0.44^a (0.096)
Wave Dummies	No	No	Yes	Yes	Yes
Category Dummies	No	No	Yes	Yes	Yes
Observations	32,551	32,354	32,354	32,354	32,354
R-squared	0.120	0.090	0.253	0.065	0.632

Notes. This table is an extended version of regression coefficients shown in the main text. Superscripts a, b, c indicate significance at 1%, 5% and 10%, respectively.

Table 5. OLS and IV Regression of Simple Logit (All Apps)

	OLS (1)	IV (2)	IV (3)	First Stage (3) Price (x2)	First Stage (3) Privacy (x3)
8.cat-enc			-1.39[a]	0.21[a]	0.61[a]
			(0.054)	(0.044)	(0.034)
9.cat-enc			-1.42[a]	-0.046	0.36[a]
			(0.054)	(0.047)	(0.036)
10.cat-enc			-1.54[a]	0.0092	0.13[a]
			(0.046)	(0.040)	(0.030)
11.cat-enc			-2.31[a]	-0.065	0.22[a]
			(0.066)	(0.058)	(0.044)
12.cat-enc			-2.21[a]	-0.054	0.051
			(0.065)	(0.057)	(0.043)
13.cat-enc			-2.13[a]	-0.13	0.26[a]
			(0.10)	(0.091)	(0.069)
14.cat-enc			-1.82[a]	0.035	0.15[a]
			(0.053)	(0.046)	(0.035)
15.cat-enc			-1.21[a]	0.14[b]	-0.055
			(0.068)	(0.059)	(0.045)
16.cat-enc			-2.29[a]	0.14	0.48[a]
			(0.11)	(0.094)	(0.071)
17.cat-enc			-0.83[a]	0.13[a]	0.038
			(0.057)	(0.050)	(0.038)
18.cat-enc			-2.50[a]	0.028	-0.0082
			(0.069)	(0.061)	(0.046)
19.cat-enc			-1.22[a]	0.050	0.16[a]
			(0.047)	(0.041)	(0.031)
20.cat-enc			-0.68[a]	0.26[a]	0.20[a]
			(0.076)	(0.066)	(0.050)
21.cat-enc			-1.70[a]	0.13[a]	0.24[a]
			(0.054)	(0.047)	(0.035)
22.cat-enc			-0.089	-0.094	-0.0072
			(0.098)	(0.086)	(0.065)
23.cat-enc			-1.41[a]	-0.13	-0.014
			(0.093)	(0.081)	(0.062)
24.cat-enc			-2.34[a]	-0.056	0.30[a]
			(0.070)	(0.062)	(0.047)
25.cat-enc			-2.37[a]	0.21[a]	0.0025
			(0.066)	(0.058)	(0.044)
26.cat-enc			-0.69[a]	0.075	0.17[a]
			(0.088)	(0.077)	(0.059)
27.cat-enc			-0.97[a]	0.096[a]	0.057[b]
			(0.043)	(0.037)	(0.028)
Wave Dummies	No	No	Yes	Yes	Yes
Category Dummies	No	No	Yes	Yes	Yes
Observations	32,551	32,354	32,354	32,354	32,354
R-squared	0.120	0.090	0.253	0.065	0.632

Notes. This table is an extended version of regression coefficients shown in the main text. Superscripts a, b, c indicate significance at 1%, 5% and 10%, respectively.

Table 5. OLS and IV Regression of Simple Logit (All Apps)

	OLS (1)	IV (2)	IV (3)	First Stage (3) Price (x2)	First Stage (3) Privacy (x3)
28.cat-enc			-1.88[a]	-0.033	0.11[c]
			(0.085)	(0.075)	(0.057)
29.cat-enc			-2.13[a]	0.24[a]	0.14[a]
			(0.059)	(0.050)	(0.038)
30.cat-enc			-2.54[a]	-0.020	-0.070
			(0.12)	(0.10)	(0.079)
Wave Dummies	No	No	Yes	Yes	Yes
Category Dummies	No	No	Yes	Yes	Yes
Observations	32,551	32,354	32,354	32,354	32,354
R-squared	0.120	0.090	0.253	0.065	0.632

Notes. This table is an extended version of regression coefficients shown in the main text. Superscripts a, b, c indicate significance at 1%, 5% and 10%, respectively.