ESTIMATING DEMAND FOR ZERO PRICED PRODUCTS AND THE VALUE OF PERSONAL DATA

Farasat A.S. Bokhari[†] Michael E. Kummer[‡]

[†]School of Economics and Centre for Competition Policy,
University of East Anglia

http://www.uea.ac.uk/economics
http://competitionpolicy.ac.uk

Program on Economics & Privacy
Research Roundtable for Privacy Fellows
University of Arizona, Jan 16-18, 2020

Motivation and Agenda

University of East Anglia — Centre for Competition Policy

- Value of personal information in apps
  - Apps often collect personal information from end users
  - Some of it is not necessary for functioning of the app
  - But also allows the developer to monetize their app
    - 20 percent of apps in Google Play Store are paid apps
    - 80 percent rely on in-app advertising, in-app purchases, and data trade
    - data sharing for monetization purposes is common
- How do consumers value data?
  - Increasing the cost of anonymity can benefit consumers, but only up to a point, after which the effect is reversed (Taylor et al., 2010)
  - Some concerns that reduced privacy is disadvantageous for consumers
  - How much (dis) utility is associated with giving up private information
- Concentration and Information
  - Higher concentration raises concerns about prices and consumer welfare
  - Does the same apply to privacy and information?
  - Will a merger between app developers increase the level of information extracted from end users?
  - What is the welfare implication when products are priced at zero?

Motivation and Agenda

- Agenda: 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
- Standard merger analysis ....
  - estimate demand parameters for a well defined market
  - assume a model of competition where firms strategically set prices and level of information extraction (an index)
  - predict post-merger prices and value of index
  - estimate implied change in consumer welfare
  - evaluate other policy changes such as restricting firms access to user data or increasing consumer protection

METHODOLOGY

- Utility for individual n from downloading app j is

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

  where y_n is the income and p_j, i_j and x_j are the price, information index and other characteristics of app j; terms ξ_j and ϵ_{nj} are the unobserved (to the econometrician) product characteristics and the idiosyncratic error term respectively
- We propose to estimate the parameters via a logit (or random coefficient logit) 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
- Firms set p_j and i_j to maximize profits which in turn depend on number of downloads

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

  - need to develop a model of competition that predicts p^* and i^* with changes in market structure (or with other policy changes)
- Compute change in consumer surplus due to merger/policy change that effects p^* and i^*

$$\Delta E(CS_n) = \frac{1}{\alpha}\left[\ln\left(\sum_j e^{V^1_{nj}}\right) - \ln\left(\sum_j e^{V^0_{nj}}\right)\right]$$

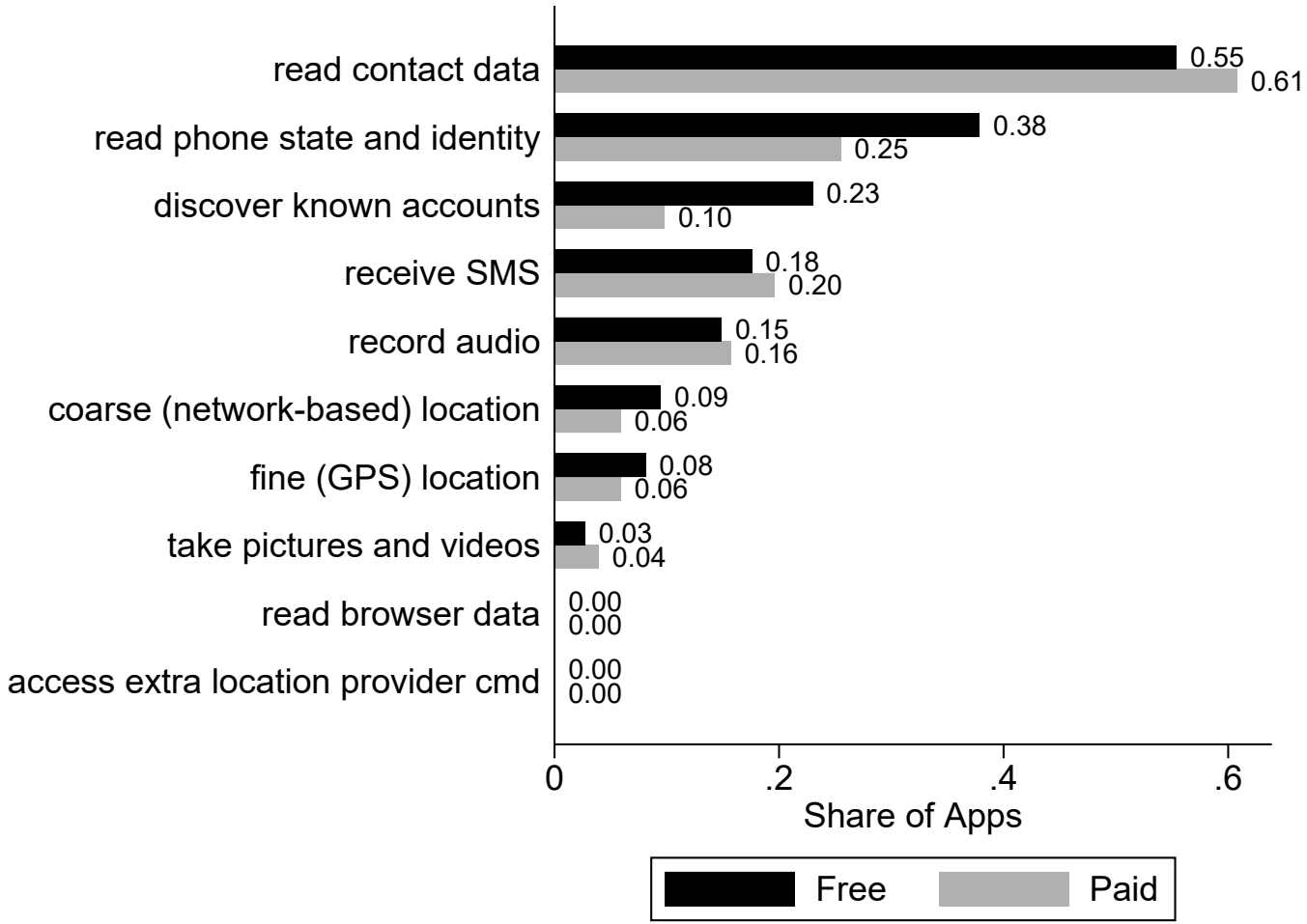
IMPLEMENTATION

- Google Play Store offered approximately 400,000 apps for download in various categories such as 'Health', 'Weather', 'Action', 'Music', 'Communication' etc.); we will use data previously used in Kummer and Schulte (2019)
  - Information is available at a monthly frequency for 2012 and includes data on downloads, price, level of privacy setting and other app characteristics
  - Quantity measures exists in a discrete form as number of new installations as well as new ratings (the two demand measures are highly correlated)
- Selection criteria – app group must have
  - variation (cross-sectional, i.e., across individual products) in price p_j and information i_j
  - exogenous source of variation in in order to be able to estimate model parameters
  - outside options s_0 varies over markets
- Email clients
  - android requires Google account and comes with Gmail client; outside option is Gmail or browser
  - changes in total number of android devices or installed base of gmail gives a measure of the size of the market Q_t; share is just q_{jt}/Q_t; outside share is $s_{0t} = 1 - \sum_j^J s_{jt}$
  - Instrument for price and information; (1) size of app and (2) developers propensity monetize information as a business model based on average value of other apps

IMPLEMENTATION
DISTRIBUTION OF MAIN VARIABLES (EMAIL CLEINTS)

	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
$D\_MTurkEP2$	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

Implementation
Prevalence of permissions (email clients)



- The standard logit model needs to be adjusted:

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

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

- For email clients gmail is the outside option on Android, and hence we estimate:

$$\ln(s_{jt}) - \ln(s_{0t}) = (\mathbf{x_{jt}} - \mathbf{x_{0t}})\beta - \alpha(p_{jt} - p_{0t}) + (\xi_{jt} - \xi_{0t}). \qquad (0.1)$$

Demand Measure
Categorial installations vs. ratings

- Our Measurement of Demand is imperfect.
- Option 1: Categorial Installations: (50,000-100,000; 100,000-500,000 etc.)
  - Installations are generally preferable.
  - However, lack of precision
  - No visible variation across time!
- Ratings:
  - Ratings are not Installations
  - However, highly correlated.
  - Precise measure with more variation over time!

→ Our preferred measure are Ratings.

ENDOGENEITY
PRICE AND PRIVACY CHOICES ARE PART OF THE MODEL

- We model price and privacy as profit-maximizing choice.
- This implies that they are determined inside the model

→ We cannot simply take them as given and regress market share on price/privacy.

- We use Instrumental Variables, and have access to 3 sets of IVs:
  - cost shifters (code size): direct predictor of price.
  - competitors' characteristics z_comp:
    - predictive, but really exogeneous? not ideal.
  - developers behavior on other apps (in other categories): "devcat"
    - e.g.: permissions required in other apps, or price charged elsewhere.

RESULTS
FIRST RESULTS (EMAIL CLIENTS)

	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) 8)	(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)
z_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

- OLS (1) cannot be trusted. Suggests that users prefer intrusive apps.
- IV(2) uses avg. price and avg privacy - weak IV, IV(3) uses avg. size and avg. privacy.
- First stages (4) highlight that size based IV is more relevant.

RESULTS

INTERPRETATION OF MAIL RESULTS:

(1) To be taken with a grain of salt.

(2) Very much variability.

ISSUES:

(1) Robustness?

- For the narrow category, results are not very robust.
- Potential issue of multicollinearity in the instruments.
- Large and insignificant coefficients.

(2) Loss of Observations

- Starting from 651 observations:
- 130 are lost when computing installation/ratings growth.
- over 220 have 0 market share and are not used in the logit.
- less than 50% of the observations are used.

Discussion and Outlook

TAKING STOCK:

- We have implemented a first conditional logit estimator, based on email-clients.
- We use industry standard (gmail) as non-zero outside option and implement IV approach.

LIMITATIONS:

- The approach suffers from serious limitations, most specifically:
  - Challenges to find predictive instruments
  - Long tail with "zero" market share:

  $\rightarrow$ Standard methodology induces a serious loss of observations.

WAY FORWARD? - We are exploring two improvements!

(1) Attempt a broader analysis across categories.

(2) Leverage the information in a 0 market share.

[ALTERNATIVE 1: Use all Data]

Option 1: Broadening the Scope
Using all categs

[Descriptives Main Dataset (all categs)]

ESTIMATION
IMPLEMENTATION OF OUTSIDE OPTION/MARKET SIZE

- In an analysis across categories, the outside good becomes,
  - "not using any app from category j."
- The Market Size is defined as
  - new Android Adopters + Old Android Users that did not use apps from category up until $t-1$
- Old Android users are computed based on
  - All Users in previous month - Category Adopters in previous month (total category installations)

OLS and IV Regression

Initial Result - 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 (0.0068)	-0.33^a (0.049)	-0.16^a (0.047)		
x_3 - privacy	-0.042^a (0.0081)	-0.13^a (0.020)	-0.066^a (0.018)		
x_5 - # of clean permissions	0.13^a (0.0045)	0.16^a (0.0075)	0.12^a (0.0068)	0.0011 (0.0032)	0.26^a (0.0024)
x_6 - average ratings	0.29^a (0.014)	0.29^a (0.014)	0.32^a (0.013)	0.019 (0.012)	-0.099^a (0.0088)
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} - Video (1/0 dummy, 1 if video	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)
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	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	.	.	0.065	0.632

All regressions include dummies for age ratings of an app. Superscripts a, b, c indicate significance at 1%, 5% and 10%.

- Col 1 OLS, Cols 4 and 5: First Stage Regressions.
- Col 2 and 3: 2 alternative IVs (with and wo category dummies)

- Main Instruments are:
  - price: developer's average price on other apps (in other categories, where possible).
  - privacy: developer's average number of permissions on other apps (in other catgoeries, where possible)
  - these are highly predicitive of the endogeneous variables.
- Privacy coefficient between -0.13 and -.066, diminishes by 50% when controlling for category and wave.
- Price analogously diminshes by 50%

- Coefficients seem more reasonably bounded now.

[ALTERNATIVE 2: Use the information in 0 market shares?]

OPTION 2: USE UNUSED APPS
LEVERAGING THE INFORMATION CONTAINED IN 0-DEMAND

OBSERVATION:

- There is information in a 0 market share
- 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

FUTURE WORK: How to leverage this insight systematically?

- Add an analysis of sample selection
- Find other ways of considering the information about unused apps

CONCLUSION
SUMMARY AND ONWARDS

TAKING STOCK:

- We have implemented a first conditional logit estimator, based on email-clients.
- We use industry standard (gmail) as non-zero outside option.
- We implement several IV approaches and apply them in estimations.

LIMITATIONS:

- The approach suffers from serious limitations, most specifically:
  - Challenges to find predictive instruments
  - Long tail with "zero" market share:

  $\rightarrow$ Standard methodology induces a serious loss of observations.

WAY FORWARD? - We will explore two improvements.

- First, attempt a broader analysis across categories.
- Second, there is information in a 0 market share – how to leverage it?
- And further: simulate merger or other policy changes etc.

REFERENCES

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