// NO.19-064 | 12/2019

DISCUSSION PAPER

// REINHOLD KESLER, MICHAEL E. KUMMER, AND PATRICK SCHULTE

Competition and Privacy in Online Markets: Evidence from the Mobile App Industry

Leibniz Association

ZEW

Competition and Privacy in Online Markets:[*]

Evidence from the Mobile App Industry

Reinhold Kesler[†] Michael E. Kummer[‡] Patrick Schulte[§]

December 4, 2019

Work in Progress

Abstract

Policy makers are increasingly concerned about the combination of market power and massive data collection in digital markets. This concern is fueled by the theoretical prediction that more market power causes firms to collect ever more data from their users. We investigate the relationship between market power and data collection empirically. We analyze data about more than 1.5 million mobile applications in several thousand submarkets of Google's Play Store. We observe these data for over two years and combine information on an app's data collection with information about its competitive environment. Our analysis highlights a robust positive relationship between market power and data collection. We find that more data are being collected in concentrated markets, and apps with higher market shares collect more data. This pattern robustly emerges across a series of cross-sectional and panel regressions as well as a series of specifications that exploit exogenous variation.

Keywords: Competition, Market Power, Privacy, User Data, Apps

JEL Classification Numbers: L17, D4, D85, D29

[*] Financial support from the Thyssen foundation is gratefully acknowledged. The paper represents the authors' personal opinions and does not necessarily reflect the views of the Deutsche Bundesbank or its staff.

[†] University of Zurich; Plattenstrasse 14; CH-8032 Zurich, Switzerland; reinhold.kesler@business.uzh.ch.

[‡] University of East Anglia, Georgia Institute of Technology, and ZEW Mannheim; 2.25E Registry And Council House, Norwich NR4 7TJ, UK; M.Kummer@uea.ac.uk.

[§] Deutsche Bundesbank; Wilhelm-Epstein-Strasse 14; D-60431 Frankfurt, Germany; patrick.schulte@bundesbank.de.

1 Introduction

A defining feature of the online market for mobile applications (apps) is that the services are frequently provided at no monetary cost. Instead, app developers often request access to the users' personal data. While this practice may be a welcome bargain that allows developers to monetize a resource that users are willing to share, it also raises concerns about the privacy of users as they may often lack knowledge, bargaining power, and choice. Specifically, app developers' strategies of collecting personal user data and sharing it with outside parties in exchange for other services has raised considerable concern by policy makers and regulators, as is also shown by the recent case of the German competition authority on Facebook's data policy claiming an abuse of Facebook's dominant position through exploitative business terms by combining different sources of data without the users' consent.[1] This concern is aggravated by the fact that some of these services have considerable market power, as in the case of Facebook for the market of social networks. Arguably, if data is the new currency in the app market, it is important to understand how competitive pressure (or lack thereof) affects the data collection practices of developers.

In this paper, we document a robust empirical relationship between market concentration and intensified collection of private user data in the market for mobile apps. We use data on more than 1.5 million apps from the Google Play Store that were observed on a quarterly basis for more than two years until January 2018. The app market is especially suitable for analyzing the relationship of market concentration and data collection, as it is characterized by a tremendous potential to collect user data at a low cost. Moreover, it is possible to observe a large number of submarkets where apps compete with varying intensity.

For our analysis, we combine data on market concentration in several thousand submarkets with information on each app's data collection. To evaluate market concentration, we use data on installations and ratings as demand measures, and exploit the Play Store's category tree as well as Google's "similar app" network to determine the submarkets, thereby having additionally a more fine-grained market definition than usual. Measures of data collection are based on the information about an app's potentially

[1] See the decision by the German competition authority (last accessed August 28, 2019).

privacy-intrusive permissions that are requested upon installation from the users. We augment this data with additional information on app developers' sharing practices with third parties.

Using our data, we are the first to document a robust positive correlation between market concentration and data collection. This pattern emerges consistently across several analyses: We first provide a careful descriptive analysis of the prevalence of data collection and of how market structure varies in the Google Play Store ecosystem. Second, we use multivariate regression, fixed effects regressions, and an instrumental variable approach to study whether more data collection ensues in more concentrated markets.

All of our approaches document a robust positive correlation between market concentration and data collection, the strength of which varies with the chosen specification. Our descriptive analysis documents that data collection is quite a common phenomenon among apps and that a considerable share of markets is highly concentrated. We also document a strong positive raw correlation between market concentration and data collection. Our regression analysis confirms the descriptive findings: Given an app's market share, apps in more concentrated environments collect more user data. Moreover, apps become more aggressive in their data collection when their market share grows.

These results are robust across multiple robustness checks involving alternative measurements of key variables, extending the analysis to data sharing, and testing, whether the results are sensitive to the choice of the sample, market, and demand. Preliminary causal analyses confirm the main finding, both when using indicators of the existence and strength of network effects as an instrument for the app's competitive environment and when exploiting a recategorization as a natural experiment similar to Ershov (2018). A first exploration of the underlying mechanism showed the relationship to be more pronounced in markets that depend on data as a currency or are economically more relevant.

Our paper contributes to the scientific and public debate in several ways. First, we provide the first large-scale empirical evidence on the positive relationship between market concentration and data collection. Our evidence is based on data from a highly relevant market and we provide novel measures for key variables, a large set of robustness checks, and implement several novel attempts to identify causal effects. Second, from a managerial perspective, the paper provides empirical evidence that aims at a better understanding on the role of data for the success of a business. Finally, our main finding documents a robust positive correlation of varying strength. This result informs the ongoing policy debate about privacy and competition between antitrust authorities, policy makers and the public.

2 Related literature

A strand of literature, which is closely related to our work, specifically studies the relationship between a firm's access to (user) data and its competitive environment. The corresponding theoretical work mostly presumes that more market power comes on average with more data collection by companies (Casadesus-Masanell and Hervas-Drane, 2015; Dimakopoulos and Sudaric, 2018). In this line of research, data is assumed to be valuable for firms, be it for targeted advertisement, the implementation of user-specific pricing, or as a direct possibility for revenue generation by selling the data to third parties. At the same time, if firms collect data about users and users realize this, it reduces the product's quality from the perspective of the users by being detrimental to their privacy (see, e.g., Kummer and Schulte, 2019). This can be more broadly related to the quality provision in oligopoly, as products can be provided with inferior quality because of the firm's market power (Tirole, 1988).

Campbell et al. (2015) study data collection in the context of competition and how this is affected by a regulation to protect privacy. They show that an opt-in regulation, where users have a fixed cost of giving their consent, might create a barrier to entry for small and new firms, especially if they cater for the needs of fewer customers. In this respect, an increasing amount of articles look at the impact of the European General Data Protection Regulation or GDPR (Goldberg et al., 2019; Jia et al., 2019). However, these studies have not considered effects on the competitive environment.

Empirical evidence on the relationship between data collection and market structure is scarce. Preibusch and Bonneau (2013) use descriptive evidence for 140 websites from five Internet industries and find that websites having no major competitor collect significantly more data than those having competitors. Sabatino and Sapi (2019) study the 2009 ePrivacy Directive, a regulation strengthening users' privacy, and its impact on the revenue and profits of European e-commerce firms. They find evidence for a negative impact on large firms, but none on small firms.

To the best of our knowledge, there are no other quantitative studies providing evidence on how the firm's market power is related to its data collection behaviour. Our contribution is to provide the first large-scale empirical evidence on this relationship in a very relevant and topical online market including novel measures for both privacy and competition. Based on the existing literature, we follow the main hypothesis that more market power comes with more data collected by firms.

Another aim of this paper is to contribute to research analyzing the economics of privacy reviewed by Acquisti et al. (2016) and Brown (2016). In this literature strand, theoretical models conceptualize privacy in online markets mostly as firms using (past) information on individuals to tailor prices or advertising (Goldfarb and Tucker, 2019). A common example of such private data on individuals comprises cookies that enable tracking users and their past behavior online, thereby facilitating targeted advertisements. Several empirical studies look at the effect of restricting the use of these cookies and consider, for example, the resulting effects on exchange prices or revenues by publishers. Their estimates range from a loss of 4 percent to 65 percent in the respective value (Goldfarb and Tucker, 2011; Johnson et al., 2019; Marotta et al., 2019). The industry estimates are much closer to the upper end, where the latest study by Google is based on a random control trial that involves disabling cookies for randomly selected users and resulting in an average loss of revenue by 52 percent of publishers.[2]

Similarly, several articles look at restricting the use of private data through the lens of privacy policies (Aziz and Telang, 2015; Johnson, 2013; Tucker, 2012, 2014). They find a decrease in revenues due to targeted ads being less successful, yet, also emphasising possible benefits for users. On a related note, some studies analyse the presence of economic returns to data with ambiguous evidence (Bajari et al., 2019; Chiou and Tucker, 2017; Claussen et al., 2019; Schaefer et al., 2018).

Finally, a few selected papers analyze the role of privacy, specifically, in app markets. The existing empirical research on this topic is mostly based on experimental and survey data, which is rather focused on the demand side (see, e.g., Egelman et al., 2013; Savage and Waldman, 2015). An exception is the study by Kummer and Schulte (2019), who exploit observational app data and analyze additionally the supply side by, for example, distinguishing free and paid apps. However, competition aspects are not within the scope

[2] See the paper by Google (last accessed August 28, 2019).

of that article.

Overall, these studies confirm the value of private data for both sides of the market. However, none of these papers have analyzed the role of the competitive environment for the firm's decision to collect data and we propose to fill this gap.

3 Background: App industry

The launch of the iPhone by Apple in 2007 marks the reinvention of the phone accompanied with the slogan to be only the beginning and shortly followed by the first phone based on the Android operating system from Google. Merely twelve years after, in the beginning of 2019, Google has 2.5 billion active mobile devices running with its operating system, while Apple has 1.4 billion.[3] In both cases, touch-based smartphones make up the vast majority. This widespread adoption can be explained by the personalization of software facilitated by applications or apps that go beyond traditional functionalities of feature or basic phones. Google and Apple make up the market for mobile operation systems, where Google comprises a share of 85 percent and Apple with close to 15 percent most of the remainder.[4] Through their respective stores, the Google Play Store (formerly Android Market) and the Apple App Store, the two platforms provide primarily apps with consumers spending a total of 101 billion US dollars worldwide in 2018.[5] While Apple makes more money from its store, Google has distinctively more apps available with close to 2.6 million in April of 2019, coming from a peak of 3.5 million in 2018 concluding a massive growth phase.[6] This also illustrates the dynamic character of the industry with an extraordinary amount of entry and exit that creates considerable variation in the competitive environment. The Google Play Store mainly consists of apps that are divided into more than 40 subcategories (including 17 for games), which themselves cover several thousands of markets as we will show.

Money in the market for mobile applications can be directly made by having paid apps or enabling in-app purchases. While the former only makes up a small fraction with around five percent of the total amount of apps, the possibility for the latter has been

[3] See the reports for Google and Apple (last accessed August 28, 2019).

[4] See, for more details, IDC's device market trends (last accessed August 28, 2019).

[5] See the State of Mobile 2019 by App Annie (last accessed August 28, 2019).

[6] See, for more details, the general statistics by AppBrain (last accessed August 28, 2019).

only introduced in 2012 and is now prevalent among double the amount of paid apps.[7] Besides this, other monetization options – in the classical sense – comprise subscriptions, rewarded products, and e-commerce, but are rather negligible.

More importantly, as in other online markets, apps are mostly for free. Instead, users give access to data and are exposed to advertising, the latter constituting a major revenue channel of app developers as the collected data can help to serve targeted advertisements. However, this is mostly limited to larger app developers, as a certain size is necessary both to cover the costs in order to sell the respective information as well as to attract advertisers. In the case of smaller developers, third parties often act as an intermediary by exchanging data for targeted ads between multiple developers and advertisers. However, third parties do not only enable this kind of access, but can also provide valuable services and functionality in exchange for user data. Finally, data from users can also be directly exchanged for money. In this respect, Christl and Spiekermann (2016) give a brief overview of the data sharing behavior by apps and the accompanied data broker industry.

4 Empirical framework

If user data is valuable to developers and users, as outlined above, the user data that developers collect resemble a non-monetary price. Building on this evidence, we want to analyze whether it is positively associated with firms' market power and market concentration. Hence, we model the amount of data collected per user as a function of various proxies of apps' market power and test this relationship empirically. In this section, we describe our main dataset, the main variables of interest, provide descriptive evidence, and describe our econometric approach.

4.1 Data

To address our research questions, we use data from the Google Play Store and AppBrain to compile a unique and innovative database, which contains quarterly product-level information for the period from 2015 to 2018 on nearly all apps available in Google's Play Store (up to 2.5 million apps per wave). We retrieve this data by means of web-scraping

[7] See the statistics by AppBrain on paid apps and for apps with in-app billing (last accessed August 28, 2019).

the complete Play Store content page of each app as described in Figure 6.

The data contains detailed information about the apps, their developers, and competitors.[8] In the following, we will mostly rely on two datasets consisting of a cross section from October 2017 and a quarterly panel ranging from October 2015 to January 2018.[9]

User data: We leverage the unique feature that the Google Play Store displays all the permissions requested by an app, thereby giving us an indication about the extent to which an app has access to user data. The amount and variety of information that developers can collect about the users through their app depends very much on each single permission the app requests upon installation from the user. In total, developers can choose among over 200 different permissions. [10] However, only some of these permissions allow to specifically collect information about users' behaviour as well as preferences and can therefore be considered privacy-sensitive and of particular interest for our study. Following Kummer and Schulte (2019), we employ a definition consisting of 25 privacy-sensitive permissions. As a result, our main variable corresponds to the number of such permissions requested by an app ($\#_{DataCollection}$). In addition, we create a dummy variable which is equal to one if an app has at least one privacy-sensitive permission ($D_{DataCollection}$).

Finally, we assess the sensitivity of these measures by also employing a category-specific definition of privacy-sensitiveness accounting for permissions of apps that are necessary for their functionality and an indicator if an app uses intrusive software libraries by third parties (taken from AppBrain).[11] We define the former as the number of privacy-sensitive permissions minus the permissions needed for functionality, thereby following the rationale that, for example, a navigation app necessitates a permission to access the location. In order to classify permissions needed for functionality, we look at the permissions used by paid apps in the categories as the benchmark, which rely less likely on data. If the share of paid apps in the category using this permission is above the

[8] For a description of the variables available in the raw data, see Appendix A.1.2.

[9] Some apps drop out of our sample due to missing values in key variables. Additionally, clusters of one app are dropped and apps have to be observed in the last three waves as well as at least two times in total. However, all competition variables are computed before these drops.

[10] See Google's documentation (last accessed August 28, 2019).

[11] Third-party libraries are software components typically easing access to services or adding functionality, which are provided by an entity other than the developer. See AppBrain for an exemplary overview (last accessed August 28, 2019).

overall average for paid apps in the Google Play Store regarding this permission, then it is a category-specific permission, which is to be subtracted. We define intrusive software libraries as sharing data with outside parties in exchange for services such as the access to ad networks, social media, or app analytics, which can be considered privacy-intrusive as well.[12]

Market power: To measure apps' market power, we make use of two standard measures being market concentration and market shares. We exploit a distinctive feature of our dataset which is the availability of information on app-specific competitors. The Google Play Store provides information on a set of 'similar apps' for each app, which are selected according to their similarity in functionality.[13] For each app between 0 and 49 of such competitors are displayed in the Play Store.[14] Based on this information, we define for each app the relevant market as well as its relevant set of competitors in the following:

First, we assume that the relevant market consists of the set of similar apps from its own category, i.e., dating apps are assumed to compete only with other dating apps but not with weather or wallpaper apps. Second, we define markets by constructing a network based on the information about similar apps. In this network, each app represents a node and an edge (or link) between two apps is established if one app is listed as a similar apps of another one. For building the network, information from all waves on similar apps is used and the link is weighted by the occurrence across quarters such that frequent similar apps are taken into account more strongly. Using network analysis, we are then able to detect clusters of apps (or 'isolated app communities') that represent market segments, which we assume to be the same across the waves.[15] An exemplary cluster comprising apps related to virtual private networks is displayed in Figure 7. We consider

[12] Intrusive libraries are classified by tags ('Ad networks,' 'Social,' and 'Analytics') provided on AppBrain and by going manually through the description pages of all development libraries on AppBrain.

[13] Although Google does not disclose the criteria, we infer this from shared experiences by developers (see, for instance, an exemplary thread on Quora, last accessed August 28, 2019).

[14] The maximum number of similar apps displayed in the Google Play Store changed over time: Until December 2016 there was a limit of 24 similar apps, whereas afterwards this has increased to 49. Additionally, the actual set of apps is in many cases even bigger such that Google varies the subset of similar apps shown.

[15] We use the R package igraph and try several cluster detection algorithms such as fastgreedy, multilevel, and walktrap. For the moment, we use the multilevel algorithm (Bliese, 2006) in our analyses, because of the moderate computational effort and the resulting clusters that have been validated by manually going through some of the clusters.

this market definition as an improvement compared to common industry-level definitions as it describes more likely the relevant choice set (and market) by consumers as opposed to a heterogeneous category encompassing ten thousands of apps. In order to show this, we compare the results when employing different aggregation levels.

Having defined for each app the relevant market and its set of competitors, we can construct variables which proxy apps' market power. A first measure of app demand is the number of ratings (*Ratings*). Following Kummer and Schulte (2019), we have the number of ratings as our demand variable of interest as it is continuous in contrast to the number of installations, which is divided into discrete intervals ranging from only 0-5 to 5-10 billion and resulting in little variation from quarter to quarter especially with increasing amount of installations. As an alternative, we use the predicted number of installations (*Pred. Installs*).[16] Both proxies measure the number of customers who consider the benefits of installing an app greater than the associated loss of privacy. The two demand measures are highly correlated (see, for more details, Kummer and Schulte, 2019).[17]

Combining the information on the relevant sets of competitors of an app and an app's demand, we then construct market shares (*Market Share*) as well as the market concentration by a Herfindahl-Hirschman Index (*HHI*). Both are standard measures of market power and typically assumed to be positively correlated with a firms' ability to raise prices above marginal cost (corresponding to a mark-up).

Control variables: In order to explain apps requesting privacy-sensitive permissions, we employ various app-specific and developer-specific measures as covariates based on previous research (see Kummer and Schulte, 2019). These variables approximate the monetization strategy (presence of an app price, in-app product price, and advertisement), the disclosure of processing personal information (presence of a privacy policy), the functionality (number of "clean", i.e., remaining permissions), quality (average rating), and general characteristics (description length, content rating, age, and category). Additionally, the number of apps by the developer is accounted for.

[16] The prediction exploits the relationship between the number of ratings and the number of installations, when an app passes an installation step.

[17] The high correlation is not surprising given that users have to install an app in order to rate it.

4.2 Descriptive evidence

Before turning to the main analysis, we first provide descriptive evidence. Table 1 shows the summary statistics of the main variables of interest for both the cross section from October 2017 and for the panel.

Table 1: Summary statistics

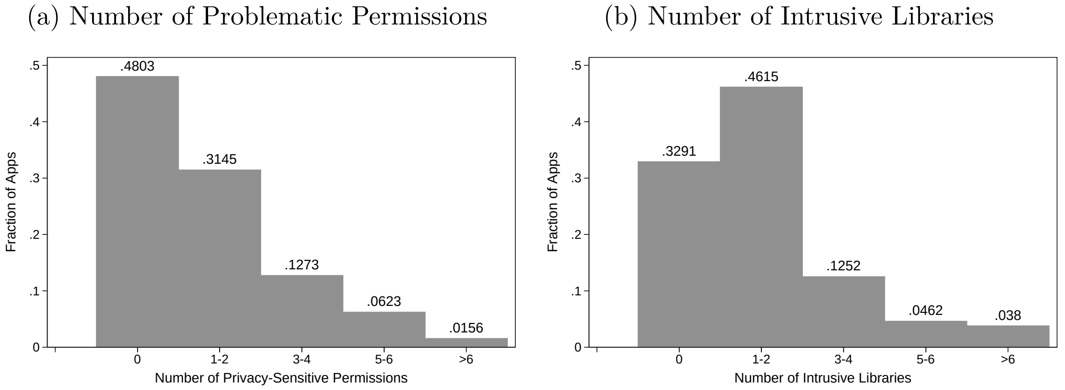
	Cross Section					Panel				
	Mean	P50	Min	Max	Count	Mean	P50	Min	Max	Count
$D_{DataCollection}$	0.52	1.00	0	1	1,336,625	0.53	1.00	0	1	11,477,730
$\#_{DataCollection}$	1.33	1.00	0	16	1,336,625	1.36	1.00	0	16	11,477,730
HHI (Ratings)	0.15	0.08	0	1	1,336,625	0.13	0.07	0	1	11,477,730
HHI (Pred. Installs)	0.17	0.10	0	1	1,336,625	0.15	0.08	0	1	11,477,730
Market Share (Ratings)	0.02	0.00	0	1	1,336,625	0.02	0.00	0	1	11,477,730
Market Share (Pred. Installs)	0.02	0.00	0	1	1,336,625	0.02	0.00	0	1	11,477,730
Log. Ratings	2.90	2.48	0	17	1,336,625	3.00	2.56	0	18	11,477,730
D_{Paid}	0.07	0.00	0	1	1,336,625	0.07	0.00	0	1	11,477,730
$\#_{CleanPerms.}$	5.22	5.00	0	174	1,336,625	5.17	5.00	0	182	11,477,730
$D_{InAppProduct}$	0.09	0.00	0	1	1,336,625	0.09	0.00	0	1	11,477,730
Avg. Rating	4.03	4.10	1	5	1,336,625	4.07	4.20	0	5	11,477,730
Length Description	1,065.01	730.00	1	11,684	1,336,625	1,048.42	727.00	1	11,697	11,477,730
$D_{IncludesAds}$	0.51	1.00	0	1	1,336,625	0.25	0.00	0	1	11,477,730
Content Rating	4.98	5.00	1	6	1,336,625	5.51	5.00	1	10	11,477,730
$D_{PrivacyProfile}$	0.36	0.00	0	1	1,336,625	0.27	0.00	0	1	11,477,730
$\#_{AppsbyDeveloper}$	89.63	8.00	1	4,690	1,336,625	77.61	8.00	1	5,171	11,477,730
App Age	1,073.48	982.00	-15	3,334	1,336,625	1,163.38	1,069.00	-15	3,334	11,477,730
$\#_{Libraries}$	4.69	3.00	0	51	333,613	4.68	3.00	0	53	2,036,135
$\#_{IntrusiveLibraries}$	1.65	1.00	0	36	333,613	1.65	1.00	0	37	2,036,135
$D_{IntrusiveLibraries}$	0.67	1.00	0	1	333,613	0.67	1.00	0	1	2,036,135

4.2.1 Collection of user data

The summary statistics show that data collection is quite a common phenomenon among apps. We observe that close to 50 percent of all apps are able to collect data through privacy-sensitive permissions. The left panel of Figure 1 in addition illustrates the distribution for the number of problematic permissions apps have access to and shows that around 20 percent of all apps have three or more of such permissions. On the other hand, about two thirds of apps use intrusive software libraries by third parties with an overall average of close to two of such libraries, while the average amount of libraries in general is close to five.[18] The right panel of Figure 1, displaying the distribution for the number of intrusive libraries, shows that 46 percent use only one or two, while the share falls rather quickly with only four percent of apps having six libraries or more.

[18] Note that, due to rate limiting, the information from AppBrain is only available for a subsample.

Figure 1: User data measures


(a) Number of Problematic Permissions

(b) Number of Intrusive Libraries

Notes: Based on the cross section from October 2017, the figures show histograms for two measures of data collection: Panel (a) shows the distribution of the number of problematic permissions of an app. Panel (b) shows the distribution of number of intrusive software libraries.

Underlying these average values, as Figure 8 in the Appendix illustrates, there is considerable cross-category heterogeneity. In some app categories such as Travel & Local, Maps & Navigation, Dating, and Communication around 80 percent of all apps have at least one permission which allows them to access private user data, whereas for example in Board, Personalization, as well as Art & Design less than 30 percent of apps have such permissions. Similarly, intrusive software libraries are most prevalent among apps in game categories and Dating with shares above 80 percent, while less than a quarter of apps in the category Libraries & Demo have such a library (figures not reported).

4.2.2 Competition in the market for mobile apps

Regarding competition, we document an especially high degree of heterogeneity among the various markets in the mobile app industry (see Figure 2 for histograms of the key variables). For example, the number of competitors within markets ranges from 2 to way above 1000 in single cases of our cluster-based market definition. Similarly, there is considerable heterogeneity in market concentration ratios, with the HHI ranging from zero to one.[19] Accordingly, we find for a considerable share of markets a high concentration ratio of above 0.5 (ten percent, when looking at all clusters with more than four apps, and four percent, when looking at clusters larger than ten apps). The average app is active in a market with an HHI equal to 0.15 (averaged across clusters). Additionally,

[19] Note that we compute the HHI on a range between 0 and 1, rather than using the definition that ranges from 0 to 10,000.

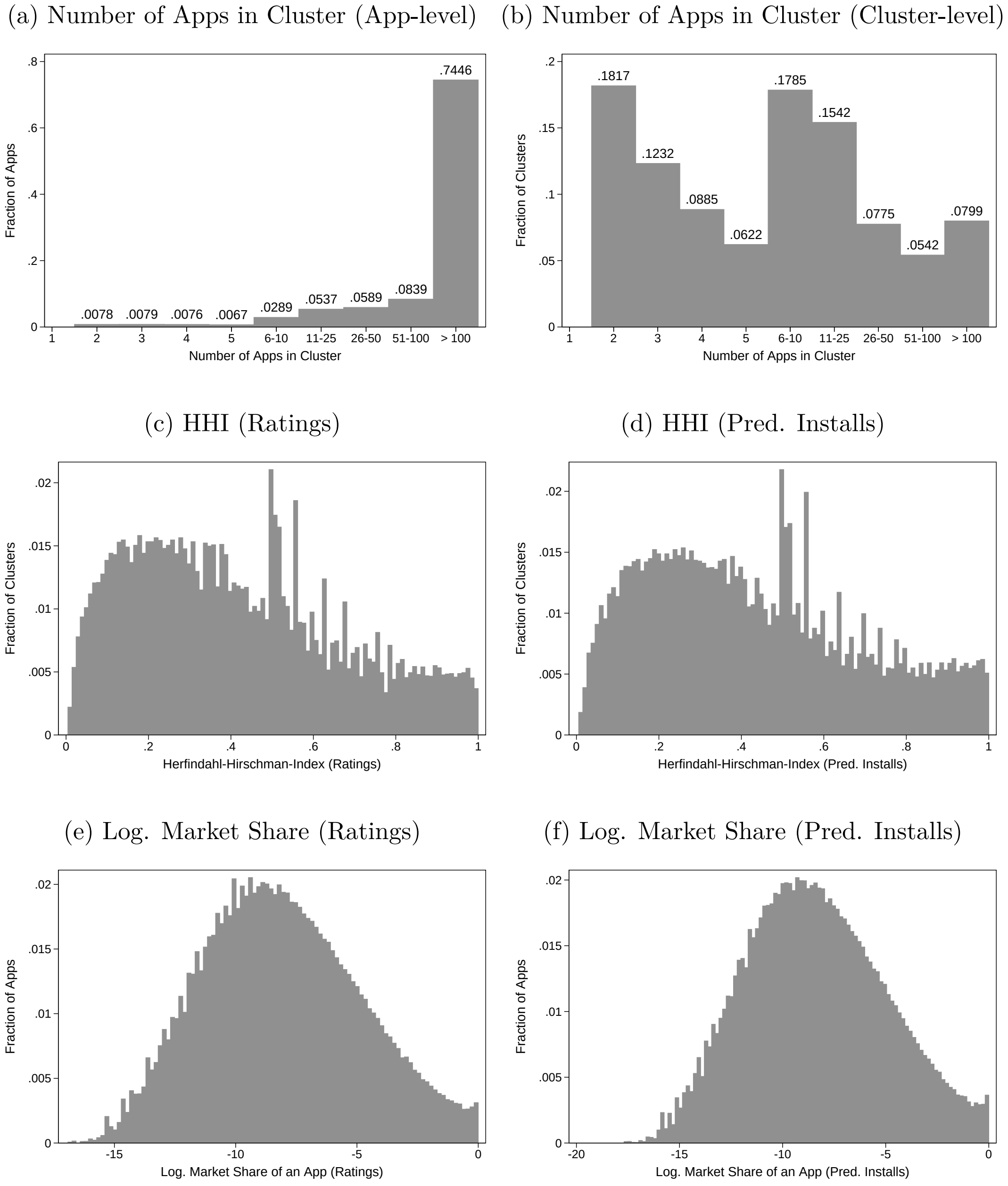
within markets, apps have very different market shares: the average app has a market share of two percent, however, the median app has a market share of below one percent indicating the long tail towards higher market shares. Similarly, there is considerable heterogeneity across categories (see Figure 9). Accordingly, the market concentration is higher for clusters in game categories and Communication, whereas it is lower in clusters of categories for Weather and Art & Design, which resonates well with the presence of network effects.

4.2.3 Relationship between data collection and competition

Figures 3 and 4 visualize the relationship between market power and data collection in our data. Figure 3 contains three bar panels which show how the HHI is related to various measures of permissions. For example, panel (a) shows that apps active in highly concentrated markets (*HHI > 0.75*) request on average around 15 percent more privacy-sensitive permissions than the overall average app, whereas in less concentrated markets (*HHI < 0.1*) apps request 2.5 percent less problematic permissions than the overall average app. The figure highlights the positive cross-sectional correlation between market concentration and apps' number of problematic permissions. In contrast, panel (b) illustrates the same relationship for unproblematic permissions, which is comparably weak, i.e., there is no clear positive or negative relationship between the HHI and the number of clean permissions and the effect sizes are also comparably small. Not surprisingly, panel (c), which shows the relationship between market concentration and the share of problematic permissions in total permissions, suggests that also the share of problematic permissions is strongly and positively correlated with the market concentration. Taken together, the figures above show a positive raw correlation between market concentration and the access to more sensitive permissions by apps. In contrast, we found no such relationship for functionality-relevant permissions, which do not allow collecting sensitive user data. The size of this surface correlation would imply a "data mark-up" of almost 20 percent (comparing markets having a very small HHI with a very high HHI).

Analogously, in Figure 4 we illustrate the relationship between apps' market share and the requested permissions. Panel (a), for example shows that apps with a very small market share (*Market Share < 0.01*) request on average 2.5 percent less problematic permissions, whereas apps with a very large market share (*Market Share > 0.3*) request

Figure 2: Competition measures


(a) Number of Apps in Cluster (App-level)

(b) Number of Apps in Cluster (Cluster-level)

(c) HHI (Ratings)

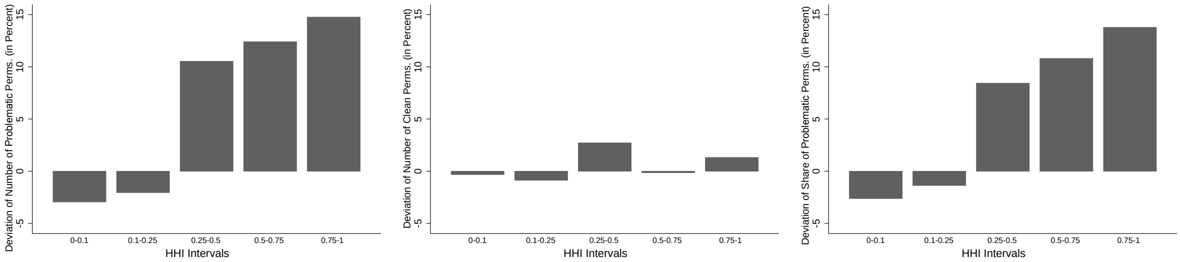
(d) HHI (Pred. Installs)

(e) Log. Market Share (Ratings)

(f) Log. Market Share (Pred. Installs)

Notes: Based on the cross section from October 2017, the figures show histograms for three competition measures. Panel (a) and (b) show the distribution of the number of apps in a cluster, both on app-level and cluster-level. Panel (c) and (d) show the distribution of the HHI, and panel (e) and (f) show the distribution of the logarithmic market share, distinguished by the two demand definitions (ratings and predicted installations), respectively.

Figure 3: Market concentration intervals and permissions



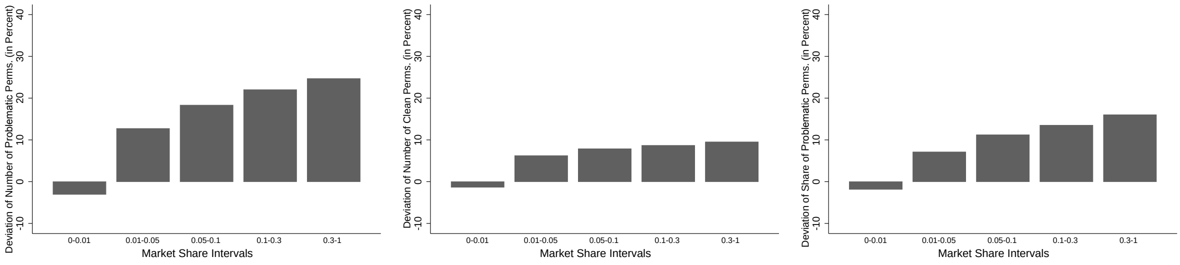
(a) Problematic

(b) Clean

(c) Share

Notes: Based on the cross section from October 2017 and the number of ratings as the demand, the figures show bar graphs for the relationship between HHI and three measures of permissions. The figures show how the average value in each HHI interval deviates (in percent) from the overall average value of the respective permission measure. Panel (a) shows it for the number of problematic permissions, panel (b) shows it for the number of unproblematic (clean) permissions, and panel (c) shows it for the share of problematic permissions in total permissions.

Figure 4: Market share intervals and permissions



(a) Problematic

(b) Clean

(c) Share

Notes: Based on the cross section from October 2017 and the number of ratings as the demand, the figures show bar graphs for the relationship between market shares and three measures of permissions. The figures show how the average value in each market share interval deviates (in percent) from the overall average value of the respective permission measure. Panel (a) shows it for the number of problematic permissions, panel (b) shows it for the number of unproblematic (clean) permissions, and panel (c) shows it for the share of problematic permissions in total permissions.

on average around 25 percent more permissions than the overall average app. Thus, similarly, for the market share there is a strong and positive relationship with apps' ability to collect user data. However, in contrast to the relationship between the HHI and requested permissions, for the market share, we find a positive relationship with clean permissions in panel (b). However, its size is small compared to the one with problematic permissions (apps with a very high market share just request less than 10 percent more clean permissions than the average app). This positive relationship for unproblematic is not surprising and might be explained through apps' functionality, which should be positively correlated with both their market share and number of unproblematic permissions. Similarly, panel (c) confirms that the share of problematic permissions in total permissions is also positively related to the app's market share. Thus, these figures illustrate that the market share of an app comes also with a more intensive use of problematic permissions and thus the ability to collect more information about users.

The descriptive results for both the market concentration and market share suggest a rather monotonic behavior, but, of course, can only give an indication, as we do not control for anything further, which is to be outlined in the following.

4.3 Econometric specification

4.3.1 Correlational analysis

In a first step, we estimate the following baseline model for our cross section of apps (from October 2017):

$$Data_i = \alpha + \beta_1 MP_i + \gamma Demand_i + \theta X_i + \epsilon_i, \tag{1}$$

where $Data_i$ is the amount of data collected per user by app i and is measured by the number of problematic permissions an app requests upon installation. Market power (MP_i) is measured via: (1) market concentration using a Herfindahl-Hirschman Index (HHI) and (2) the logarithmic market share ($Log.\ Market\ Share$).[20] A positive value for our coefficient of interest, β, would indicate that a more concentrated market is related on average with more problematic permissions per app and user. We control for apps' own demand ($Demand_i$), which is proxied by the (log) number of ratings an app has

[20] The logarithmic market share is chosen to reduce right skewness. In a robustness check, we turn to market shares.

received, as well as for a set of explanatory variables relevant for data collection outlined before (X_i).

4.3.2 Analysis with Panel Fixed Effects

In a second step, to control for time-invariant unobserved factors, which could drive the cross-sectional relationship between data collection and market power, we use our panel, which covers the period from October 2015 to January 2018 and apply a panel fixed effects approach. Hence, we extend our baseline specification by including an app fixed effect (α_i):

$$Data_{it} = \alpha_i + \beta_1 MP_{it} + \gamma Demand_{it} + \theta X_{it} + \epsilon_{it}, \tag{2}$$

where t represents a quarter. In particular, unmeasured quality and functionality of apps could be positively correlated with both the use of permissions (which can come with additional functionality) and especially a higher market share, which, if not controlled for, would result in an upward bias.

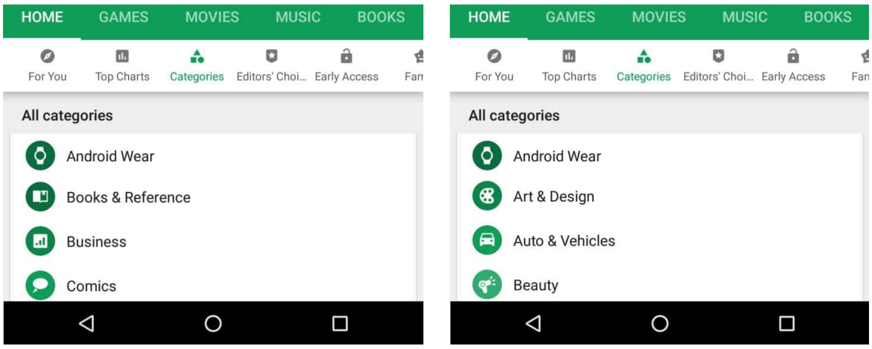
4.3.3 Endogeneity

Motivation: Although the previous methods can control for time-invariant unobserved and various time-variant observed factors, we note that it cannot account for all sources of bias. The remaining sources of endogeneity might provide biased estimation results for the true effect of market power on apps' data collection behavior: First, unobserved factors such as managerial quality or quality of the app's business plan might drive both the strategy for data collection and the app's quality (and the resulting market share). Second, data collection could also drive market shares as well as market concentration, and induce reverse causality that stands in the way of identifying the causal effect of interest. Third, our measures of market power come with considerable noise and measurement error, which will induce a classical attenuation bias. Unobserved heterogeneity and reverse causality could induce a positive bias for the estimated coefficients of both market share and market concentration. This concern is somewhat mitigated in our panel analysis, which accounts for time-invariant unobserved heterogeneity. Moreover, the bias for market concentration could be reduced because we control for market share. In contrast, a measurement error will bias our estimates towards zero, which counteracts the aforementioned upward biases.

Exploiting exogenous variation: To identify the effect of market power on data collection causally, we first utilize a recategorization of apps in the Google Play Store as a 'natural experiment.' In early September 2016, Google added eight new app categories.[21] These new categories include Art & Design, Auto & Vehicles, Beauty, Dating, Events, Food & Drink, House & Home, and Parenting.[22]

As a result, consumer search improved as the discoverability of apps that moved into these new, and rather niche, categories increased compared to before, when they were together with distinct app types in a much broader category. This is supported by surveys showing that the majority of consumers find new apps through the store, either by searching specifically or browsing through categories.[23] Figure 5 shows a screenshot of the category tree before and after the recategorization.

Figure 5: Category tree before and after the recategorization



Accordingly, we expect that competition intensified for apps in these new categories. This rationale is in line with Ershov (2018), who exploits a past recategorization of apps in game categories and finds also an increased app entry. More importantly, the announcement and the accompanied changes were not anticipated by market participants.[24] In addition, to the best of our knowledge, there was no other major policy change in the Google Play Store at the time.

We consider apps that moved into new categories as the treatment group and exclude game categories in the analysis (as they are less comparable). Admittedly, apps in existing

[21] See Google's announcement (last accessed August 28, 2019).

[22] In addition, the categories Transportation and Media & Video were renamed to Maps & Navigation and Video Players & Editors, respectively. A comparison of the categories before and after is given in Table 3.

[23] See Ershov (2018) or, for instance, Apple's statistics on search to find apps (last accessed August 28, 2019).

[24] Google claims to improve the overall search experience, while providing no particular selection criteria.

categories, where apps are moved away from are to some degree treated as well. However, as can be seen in Table 3, the apps moving away make up only a negligible fraction of the old categories. The table also shows, in which categories apps were before the change.[25] We will denote these control groups as "OLD," while we have another control group denoted by "NOT" that are apps in the remaining categories. Additionally, we run several robustness checks to analyse the sensitivity.

As an alternative approach, we utilize the exogenous variation in the intensity of network effects in an instrumental variable framework in Subsection A.3.

5 Estimation results

In the following subsection, we present our main finding: In more concentrated markets, there is an increased data collection, while apps having a higher market share also collect more data. Subsequently, we test the robustness of this finding in Subsection 5.2, which is followed by an analysis of the mechanisms that drive our main finding (Subsection 5.3).

5.1 Main result

Table 2 shows our cross-sectional and panel fixed effects analysis of the relationship between market concentration and data collection. The dependent variable in these regressions is the number of privacy-sensitive permissions that an app collects ($\#_{DataCollection}$), and we show the main result for the cross section (columns 1, 2, 5, and 6), and the full panel (columns 3, 4, 7, and 8).[26] We use two alternative demand definitions: The first definition (denoted by 1) uses the number of ratings to compute market shares and concentration. For the second definition, we predict the number of installations before computing market shares and market concentration (denoted by 2). Columns 1-4 focus on market concentration only, while columns 5-8 include also the app's market share in the specification, and the main coefficients of interest are *HHI* and *Log. Market Share.*

The results show that market concentration is positively related to data collection across all different specifications. While the relationship is stronger in the cross section, the coefficients are smaller but remain robust in the panel. The effect prevails when

[25] See Google's respective help page (last accessed August 28, 2019).

[26] A count data analysis yields qualitatively similar results. For simplification, we only show OLS results.

controlling for market share, even though market share itself is also positively related to data collection, which suggests, that the main effects may be driven by apps that gain increasing shares of the market. The number of new ratings is negatively related to data collection, which could suggest that apps collecting substantial amount of data are not as highly demanded as apps which do not request such data (in line with findings from Kummer and Schulte, 2019). Note that all these specifications control, among other things, for price, functionality (measured by the number of unproblematic permissions $\#_{CleanPerms.}$), app age, and quality (using our proxies that measure average rating and the length of the app's description). Taken together, we find a robust pattern: Apps in highly concentrated markets ($HHI > 0.8$) collect around 1-2 percent more data than apps in competitive markets ($HHI < 0.1$). In addition, players with a higher market share within a market collect more user data. We note that the estimated relationships in this specification are relatively small, but point out that the panel results might be biased towards zero because of attenuation bias.

5.2 Robustness

In this section, we analyze the robustness of our results. First, we show that our results neither critically depend on how we measure our dependent variable (data collection), nor on how we quantify our variables of interest (competition). Second, we show that our results extend to sharing the data with outside parties that offer developers functionalities and services through libraries in exchange. Third, we consider different sample, market, and demand definitions in order to rule out that the results are driven by our preferred choice. Finally, different ways to exploit exogenous variation are employed.

5.2.1 Employing alternative measures

Table 4 analyzes the relationship, when using alternative measures of data collection. As before, we show cross-sectional and panel regressions for both demand definitions. The dependent variable in columns 1-4 is a dummy variable that takes the value of one, if an app requests *any* privacy-sensitive permissions ($D_{DataCollection}$), and in columns 5-8, we use context-specific criteria based on the app categories to determine, which privacy-sensitive permissions are not necessarily functional ($\#_{Category-SpecificDataCollection}$). The effects remain robust, even when we use the category-specific measure, which accounts

Table 2: Baseline estimations

	$\#_{DataCollection}$							
	CS1	CS2	Panel1	Panel2	CS1	CS2	Panel1	Panel2
HHI	0.141***	0.126***	0.004***	0.004***	0.079***	0.079***	0.009***	0.008***
	(0.007)	(0.006)	(0.001)	(0.001)	(0.007)	(0.006)	(0.001)	(0.001)
Log. Market Share					0.013***	0.012***	0.002***	0.002***
					(0.000)	(0.000)	(0.000)	(0.000)
Log. Ratings	-0.068***	-0.068***	-0.018***	-0.018***	-0.078***	-0.079***	-0.020***	-0.020***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
D_{Paid}	-0.334***	-0.334***	0.028***	0.028***	-0.330***	-0.330***	0.028***	0.028***
	(0.004)	(0.004)	(0.009)	(0.009)	(0.004)	(0.004)	(0.009)	(0.009)
$\#_{CleanPerms.}$	0.312***	0.312***	0.267***	0.267***	0.312***	0.312***	0.267***	0.267***
	(0.001)	(0.001)	(0.002)	(0.002)	(0.001)	(0.001)	(0.002)	(0.002)
$D_{InAppProduct}$	-0.120***	-0.120***	-0.025***	-0.025***	-0.118***	-0.118***	-0.025***	-0.025***
	(0.004)	(0.004)	(0.006)	(0.006)	(0.004)	(0.004)	(0.006)	(0.006)
Avg. Rating	0.001	0.001	-0.008***	-0.008***	0.000	0.000	-0.008***	-0.008***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Log. Length Description	-0.035***	-0.035***	-0.017***	-0.017***	-0.035***	-0.035***	-0.017***	-0.017***
	(0.001)	(0.001)	(0.003)	(0.003)	(0.001)	(0.001)	(0.003)	(0.003)
$D_{IncludesAds}$	-0.324***	-0.325***	-0.046***	-0.046***	-0.321***	-0.321***	-0.046***	-0.046***
	(0.003)	(0.003)	(0.001)	(0.001)	(0.003)	(0.003)	(0.001)	(0.001)
Content Rating	0.015***	0.015***	-0.006***	-0.006***	0.013***	0.013***	-0.006***	-0.006***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
$D_{PrivacyProfile}$	0.372***	0.372***	-0.037***	-0.037***	0.373***	0.373***	-0.037***	-0.037***
	(0.003)	(0.003)	(0.001)	(0.001)	(0.003)	(0.003)	(0.001)	(0.001)
$\#_{AppsbyDeveloper}$	-0.000***	-0.000***	0.000***	0.000***	-0.000***	-0.000***	0.000***	0.000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
App Age	0.000***	0.000***	0.000	0.000	0.000***	0.000***	0.000	0.000
	(0.000)	(0.000)	(.)	(.)	(0.000)	(0.000)	(.)	(.)
Constant	-0.382***	-0.382***	0.250***	0.250***	-0.211***	-0.211***	0.268***	0.267***
	(0.013)	(0.013)	(0.020)	(0.020)	(0.014)	(0.014)	(0.020)	(0.020)
Category	Yes	Yes	No	No	Yes	Yes	No	No
Wave	No	No	Yes	Yes	No	No	Yes	Yes
Mean $\#_{DataCollection}$	1.325	1.325	1.356	1.356	1.325	1.325	1.356	1.356
Observations	1,336,625	1,336,625	11,477,730	11,477,730	1,336,625	1,336,625	11,477,730	11,477,730
Num. of Groups			1,705,215	1,705,215			1,705,215	1,705,215
Adjusted R^2	0.52	0.52	0.28	0.28	0.52	0.52	0.28	0.28

Notes: The table shows cross-sectional and panel fixed effects regressions, with the dependent variable being the number of privacy-sensitive permissions that an app collects *# DataCollection*. The coefficients of interest are *HHI* and *Log. Market Share*. Columns 1-4 focus on market concentration only, while columns 5-8 include the app's market share in the specification. Odd columns (1, 3, 5, and 7) use our ratings demand definition to compute market shares and concentration, while even columns (2, 4, 6, and 8) predict the number of installations as the demand before computing market shares and market concentration. Heteroscedasticity-robust standard errors in parentheses: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

for app-specific permission needs and negates one purpose of data collection being functionality. This strengthens our result that data is a currency in this market.

In Table 5, we employ alternative competition measures. The first two columns show how concentrated (*HHI = 0.3-0.6*) and highly concentrated (*HHI = 0.6-1.0*) markets compare to less concentrated markets (*HHI < 0.3*). In columns 3 and 4, we examine the number of competitors in the market ($\#_{NumberofCompetitors}$ in hundreds), whereas in columns 5 and 6 our concentration measure is the *Market Share of the Top 4*. Similarly, we show cross-sectional and panel regressions, however, only for the demand definition based on the number of ratings.

The results in columns 1, 2, 5, and 6 indicate a positive relationship between market concentration and data collection as before. Column 4, in contrast shows a negative relationship between the number of competitors and data collection, which is in line with the idea that less competitive markets (where less competitors are active) see higher rates of data collection. The positive correlation in the cross section might be driven by the heterogeneity between the apps. The results thus also highlight that the effect is not only driven by firms with larger market shares, but also by the number of competitors. However, a dummy variable indicating the market leader as an explanatory variable is also positively and significantly related with data collection (tables not reported). Finally, it is worth noting that the size of the effects on average data collection remains relatively small.

5.2.2 Sharing data with third parties

Table 6 shows the relationship between competition (or lack thereof) and data sharing as measured by the existence of intrusive software libraries within apps. We show cross-sectional estimates, where the dependent variable is a dummy variable equal to one if an app uses at least one intrusive software library ($D_{IntrusiveLibrary}$).[27] This dependent variable is based on our augmented data about developers' use of third-party libraries. Using these libraries implies sharing the collected data with outside parties and thus sheds light on data usage rather than collection. Our analysis shows a positive relationship, and thus reinforces the findings on data collection from before. Data sharing is more common in concentrated markets and apps with a stronger position are more likely to share their data.

5.2.3 Varying the sample, market, and demand definition

In the following analysis, we restrict the sample to a balanced panel and repeat our baseline estimations. Table 7 shows that the results persist when disregarding entry and exit observations, with the relationship being more pronounced suggesting some effect

[27] We can only show cross-sectional evidence, as the AppBrain data is only available for a few waves, making a panel difficult.

heterogeneity.

As an alternative to our fine-grained market definition, we first define markets to be equal to the categories provided by Google of which there are close to 50. This is comparable with industry-level studies and possibly more common. In a second step, we restrict our attention to the 100 most popular apps in each category regarding the number of installations, as they are more likely to be in the choice set for consumers browsing the Play Store.[28] For both of these alternative market definitions, we find in Table 8 that the positive relationship between data collection and market concentration as well as market share persists in the cross section and is larger in magnitude, while in the panel the results are less conclusive and largely insignificant. However, the results for the sample comprising the most popular apps in each category lean to those, when using our cluster-based measures.

Finally, the results remain robust, when defining the market even more narrow by taking only the similar apps of up to 49 provided for each app as well as when using the quarterly change in the number of ratings instead of the stock measure as the demand (tables not reported).

5.2.4 Exploiting exogenous variation

Before turning to the main regression analysis of the recategorization, simple graphs shall motivate the setup and show the development of apps in the treatment and control groups for key variables, respectively. Figure 10 shows the average number of privacy-sensitive permissions in new categories to decrease strongly in the periods after the recategorization (panel (a)) and the average market concentration of apps to fall as well (panel (b)), in both cases especially compared to the existing categories. This leaves us with the hypothesis that the recategorization led to a stronger competition in the respective markets, which we exploit in the following analysis. However, the summary statistics in Table 9 reveals distinctive differences between new and existing categories necessitating a careful choice of a control group.

In principle, we repeat our previous estimations and include variables indicating the recategorized apps as the treatment group and compute interactions with time dummy variables, labelled around the time of the category change. We vary the specification,

[28] The results are similar, when looking at the top 500 or top 1000 apps in each category.

control group, and sample in the following. However, all results suggest that subsequent to the recategorization, treated apps request significantly less privacy-sensitive permissions compared to the respective control group. Accordingly, Table 10 and 11 show the results of the pooled OLS and app fixed effect estimations, respectively. As before, the dependent variable is the number of privacy-sensitive permissions that an app requests. The coefficients of interest are the interactions between *Recat*, a dummy equal to one for apps that moved into new categories, and time dummy variables denoted by t_{Recat}. The column titles denote the respective sample and control group. Recategorized apps request less privacy-sensitive user data following the shock, either instantly (Table 11) or beginning with the second period afterwards (Table 10). This is robust to using different control groups such as all apps (excluding game categories), existing categories having apps that move, existing categories having no apps that move, and finally restricting to markets without the presence of a mover in accordance with our cluster-based definition (columns 1, 2, 3, and 5). Finally, limiting the attention to a balanced panel still gives the main result, though with weaker significance (column 4).

In an alternative approach motivated in Subsection A.3.1, the intensity of network effects is used as an instrument for the competitive environment of an app. Table 15 shows the main finding to be confirmed qualitatively. The results are explained in more detail in Subsection A.3.2.

5.3 Mechanism

5.3.1 Contrasting free and paid apps

Table 12 aims at shedding light on whether the relationship is more pronounced in markets that are dependent on data as a currency. Consequently, columns 1, 2, 5, and 6 analyze free apps, while columns 3, 4, 7, and 8 show the baseline estimations for paid apps, thereby differentiating between our two demand definitions, respectively. The results show that the effects are driven by free apps. This seems quite plausible, as free apps have to use data as a means for revenue generation in the form of ads, for example. Moreover, it is reassuring that paid apps refrain from collecting data even in concentrated markets, suggesting that for these apps the monetary price is still the currency instead of data.

5.3.2 Size and importance of markets

In a further step, we restrict the attention to markets that are economically more relevant, where we expect the relationship to be prevalent. For this, we restrict the sample in Table 13 to either submarkets that consist of at least ten apps throughout the observation period (columns 1-4), or have a minimum of 100,000 total installations at the beginning (columns 5-8). As hypothesized, the relationship is more pronounced in markets that are larger in size as well as in terms of demand.

5.3.3 Nonlinearity of market shares

In Table 14, we refrain from taking the logarithmic transformation of market shares and vary the functional form. Besides looking at linear market shares, we study the possibility of a non-linear relationship by including the squared market share, which has been studied in the context of competition and innovation (see e.g., Aghion et al., 2005; Cohen, 2010). The results confirm a robust relationship between market concentration and data collection, when accounting for different functional forms of the market share. Interestingly, the coefficient for the market share is only positive and significant for the cross section and insignificant for the panel, when including the linear term. Adding the squared term, the linear market share becomes positive and significant in the panel regressions as well, while the squared term is negative and significant throughout the specifications. This suggests an inverted U-shaped form, where apps with a higher market share request more privacy-sensitive permissions but only up to a point of reversal.

6 Conclusion

Market concentration in data-driven markets is a major concern of policy makers and regulators. As the digital age reduces the cost of collecting information, it is argued that user data has become the new currency of many digital goods and services. Indeed, theoretical research and descriptive evidence suggest that data collection (and the associated loss of privacy) increases with a firm's market power, similar to a price.

In this paper, we provide the first large-scale empirical evidence on the relationship of market concentration and the collection of privacy-sensitive user data. We study the market structure of the Google Play Store comprising more than 1.5 million apps in several thousand submarkets over the period from October 2015 to January 2018 on a quarterly basis and combine this data with information on each app's data collection practices as well as additional information on data sharing with third parties.

We find robust evidence that apps in markets with higher concentration collect more privacy-sensitive user data. Within such markets, apps with higher market shares collect more data. The positive relationship is robust to using alternative measurements of key variables, extending the analysis to data sharing, and varying the sample, market, and demand definition. We address endogeneity concerns by instrumenting the app's competitive environment by measures of network effects and exploit a recategorization. The main result emerges consistently across all our specifications, but the size and economic significance of the effects vary with the specification and the focal market. The relationship is more pronounced in economically relevant markets and in markets that rely on data, but further research is needed to provide a more precise quantification of the effect.

Nevertheless, our work suffers from several limitations, which we seek to overcome in ongoing research. We measure our effect of interest for the average app across all markets, but additional research shall analyze markets with different oligopolistic structures separately and study the role of entry and exit more carefully. This shall also help to further study the underlying mechanisms. In a similar vein, regressions on the market-level, accounting for a possible lag structure, and varying the functional form of variables approximating competition are to be revisited. Additionally, we plan to improve our causal analysis by refining the control group for the recategorization and the accompanied analyses.

These limitations notwithstanding, our research provides important empirical evidence about the relationship of market concentration and data collection, thereby contributing to the scientific, managerial, and policy perspective. First, it complements previous research by studying a highly relevant market at large and employing novel measures, as well as a series of analyses including the identification of causal effects. Second, it sheds light on the role of data for firm success. Third, it contributes to the recently emerged interest by competition authorities on the role of privacy for antitrust.

Our results suggest that data collection is more prevalent in concentrated markets. This finding raises questions regarding data sharing and data portability. However, even though our evidence of a positive relationship between market concentration and data collection is highly significant and robust, our estimates of its strength are sensitive to the chosen specification. Especially, the fixed effects estimation reveals the relationship to be rather small. According to our preferred specification, apps in highly concentrated markets collect on average around one to three percent more data, but further research is needed to test and validate these findings.

References

Acquisti, Alessandro, Curtis Taylor, and Liad Wagman, "The Economics of Privacy," *Journal of Economic Literature*, 2016, *54* (2), 442–92.

Aghion, Philippe, Nicholas Bloom, Richard Blundell, Rachel Griffith, and Peter Howitt, "Competition and Innovation: An Inverted-U Relationship," *Quarterly Journal of Economics*, 2005, *120* (2), 701–728.

Aziz, Arslan and Rahul Telang, "What is a Cookie Worth?," *Heinz College Working Paper*, 2015.

Bajari, Patrick, Victor Chernozhukov, Ali Hortaçsu, and Junichi Suzuki, "The Impact of Big Data on Firm Performance: An Empirical Investigation," *AEA Papers and Proceedings*, 2019, *109*, 33–37.

Bliese, Paul, "Multilevel Modeling in R (2.2) – A Brief Introduction to R, The Multilevel Package and the nlme Package," 2006.

Brown, Ian, "The Economics of Privacy, Data Protection and Surveillance," in "Handbook on the Economics of the Internet," Edward Elgar Publishing, 2016, pp. 247–261.

Campbell, James, Avi Goldfarb, and Catherine Tucker, "Privacy Regulation and Market Structure," *Journal of Economics & Management Strategy*, 2015, *24* (1), 47–73.

Casadesus-Masanell, Ramon and Andres Hervas-Drane, "Competing With Privacy," *Management Science*, 2015, *61* (1), 229–246.

Chiou, Lesley and Catherine Tucker, "Search Engines and Data Retention: Implications for Privacy and Antitrust," *NBER Working Paper No. 23815*, 2017.

Christl, Wolfie and Sarah Spiekermann, *Networks of Control – A Report on Corporate Surveillance, Digital Tracking, Big Data & Privacy*, Facultas, Vienna, 2016.

Claussen, Jörg, Christian Peukert, and Ananya Sen, "The Editor vs. the Algorithm: Targeting, Data and Externalities in Online News," *Working Paper*, 2019.

Cohen, Wesley M., "Fifty Years of Empirical Studies of Innovative Activity and Performance," in Bronwyn H. Hall and Nathan Rosenberg, eds., *Handbook of the Economics of Innovation*, Vol. 1 of *Handbook of the Economics of Innovation*, North-Holland, 2010, pp. 129–213.

Dimakopoulos, Philipp D. and Slobodan Sudaric, "Privacy and Platform Competition," *International Journal of Industrial Organization*, 2018, *61*, 686–713.

Egelman, Serge, Adrienne Porter Felt, and David Wagner, "Choice Architecture and Smartphone Privacy: There is a Price for That," in "The Economics of Information Security and Privacy," Springer, 2013, pp. 211–236.

Ershov, Daniel, "The Effects of Consumer Search Costs on Entry and Quality in the Mobile App Market," *Working Paper*, 2018.

Goldberg, Samuel, Garrett Johnson, and Scott Shriver, "Regulating Privacy Online: The Early Impact of the GDPR on European Web Traffic & E-Commerce Outcomes," *Working Paper*, 2019.

Goldfarb, Avi and Catherine Tucker, "Privacy Regulation and Online Advertising," *Management Science*, 2011, *57* (1), 57–71.

_ **and** _ , "Digital Economics," *Journal of Economic Literature*, 2019, *57* (1), 3–43.

Jia, Jian, Ginger Zhe Jin, and Liad Wagman, "GDPR and the Localness of Venture Investment," *Working Paper*, 2019.

Johnson, Garrett A., "The Impact of Privacy Policy on the Auction Market for Online Display Advertising," Technical Report 2013.

Johnson, Garrett, Scott Shriver, and Shaoyin Du, "Consumer Privacy Choice in Online Advertising: Who Opts Out and at What Cost to Industry?," *Marketing Science (Forthcoming)*, 2019.

Kummer, Michael E. and Patrick Schulte, "When Private Information Settles the Bill: Money and Privacy in Google's Market for Smartphone Applications," *Management Science (Forthcoming)*, 2019.

Marotta, Veronica, Vibhanshu Abhishek, and Alessandro Acquisti, "Online Tracking and Publishers' Revenues: An Empirical Analysis," *Working Paper*, 2019.

Preibusch, Soeren and Joseph Bonneau, "The Privacy Landscape: Product Differentiation on Data Collection," in "Economics of Information Security and Privacy III," Springer, 2013, pp. 263–283.

Sabatino, Lorien and Geza Sapi, "Online Privacy and Market Structure: Theory and Evidence," *DICE Discussion Paper 308*, 2019.

Savage, Scott J. and Donald M. Waldman, "Privacy Tradeoffs in Smartphone Applications," *Economics Letters*, 2015, *137*, 171–175.

Schaefer, Maximilian, Geza Sapi, and Szabolcs Lorincz, "The Effect of Big Data on Recommendation Quality: The Example of Internet Search," *DIW Berlin Discussion Paper 1730*, 2018.

Tirole, Jean, *The Theory of Industrial Organization*, MIT Press, 1988.

Tucker, Catherine, "The Economics of Advertising and Privacy," *International Journal of Industrial Organization*, 2012, *30* (3), 326–329.

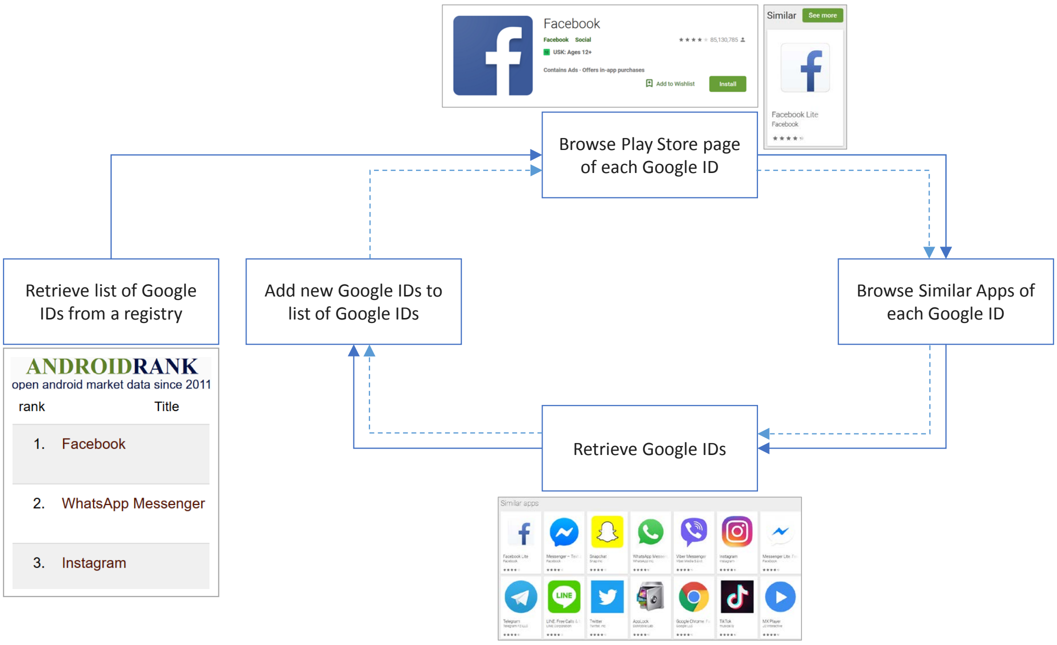
_ , "Social Networks, Personalized Advertising and Privacy Controls," *Journal of Marketing Research*, 2014, *51* (5), 546–562.

A Appendix

A.1 Additional information on dataset

A.1.1 Collection of data

Figure 6: Web-scraping of the Google Play Store


Facebook
Facebook Social
USK: Ages 12+
Contains Ads · Offers in-app purchases
Add to Wishlist
Install
Similar
See more
Facebook Lite
Facebook
Browse Play Store page of each Google ID
Retrieve list of Google IDs from a registry
Add new Google IDs to list of Google IDs
Browse Similar Apps of each Google ID
ANDROIDRANK
open android market data since 2011
rank
Title
1. Facebook
2. WhatsApp Messenger
3. Instagram
Retrieve Google IDs
Similar apps

Notes: The figure shows the process of retrieving apps in the Google Play Store by means of web-scraping. Starting with a registry of the most relevant apps, e.g., from AndroidRank, new apps are identified by the similar apps provided on the Play Store page for each app.

A.1.2 Description of data

The raw data include the following app-specific information which we use to construct our measures of competition, and developers' data access as well as our control variables:

Competitor information:

- names and IDs of similar apps.

User data (or privacy):

- all permissions that apps are requesting (upon installation) and that apps require to perform certain functions (in total more than 200 permissions, including, e.g., 'network access,' 'read contents of USB,' 'read contact data,' 'read browser data,' 'read sensitive log data'),

- additional information about these permissions (special flag by Google, considered privacy-sensitive by researchers, etc.),
- name and type of libraries (retrieved from AppBrain).

Control variables:

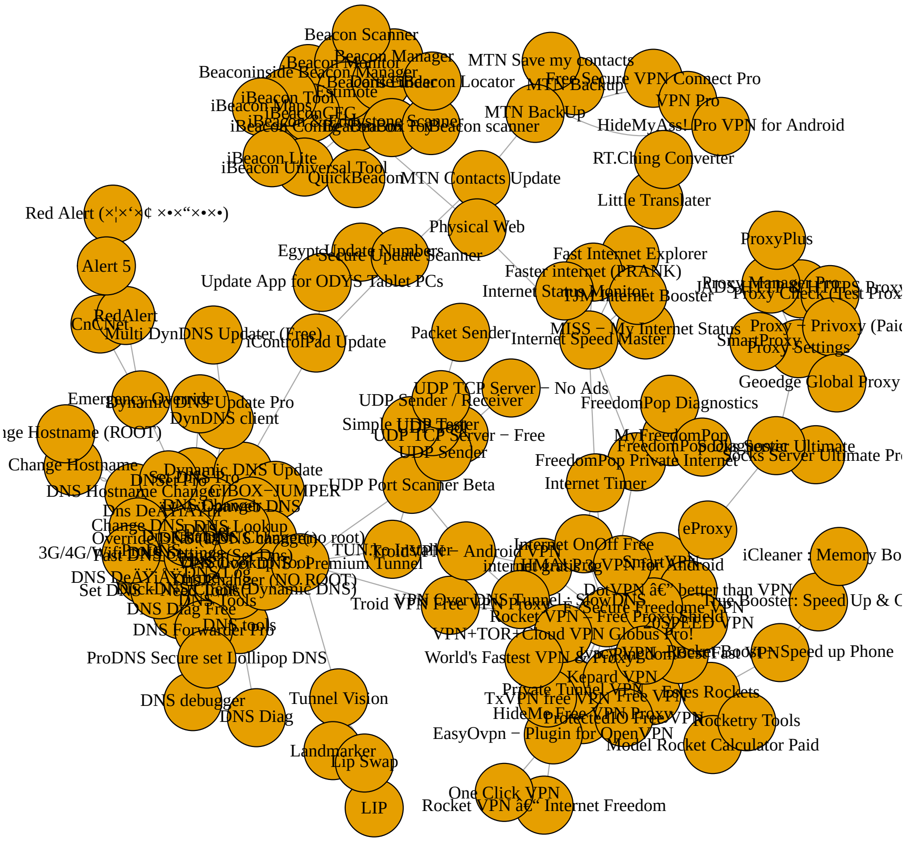
- total number of installations of an app,
- number and values of quantitative ratings (from 1 to 5 stars),
- information on updates: date, textual information on what is new, and version number,
- price (in Euro),
- existence of in-app purchases and the price range of such items in Euro,
- existence of in-app advertisements,
- app category (e.g., Racing, Personalization, Traveling, Weather, Social, Health & Fitness, Finance, Communication, etc.),
- apps' description (length and content) and its illustration in the Play Store (video and screenshot availability),
- app age (retrieved from AppBrain),
- content rating (USK) and 'may contain' warnings,
- availability of interactive elements (e.g., 'users interact,' 'digital purchases,' etc.),
- is the app an editor's choice (yes/no),
- Android version required for installation,
- presence of a privacy policy,
- contact information of the app (including website, e-mail, and address).

Developer-specific information:

- name of the developer,
- top developer status (yes/no),
- number of its apps,
- set of its available apps.

A.1.3 Market definition based on clusters

Figure 7: Exemplary cluster of similar apps



Notes: The figure shows the cluster of apps that allow users to access the Internet through a virtual private network. A node in the network represents an app, and an edge means that one app is the other one's set of "similar apps." While all the apps in this figure belong to the same cluster or submarket in our analysis, some apps are clearly closer in the "similar app" network than others.

A.2 Tables and figures

Figure 8: Share of apps with problematic permissions by category

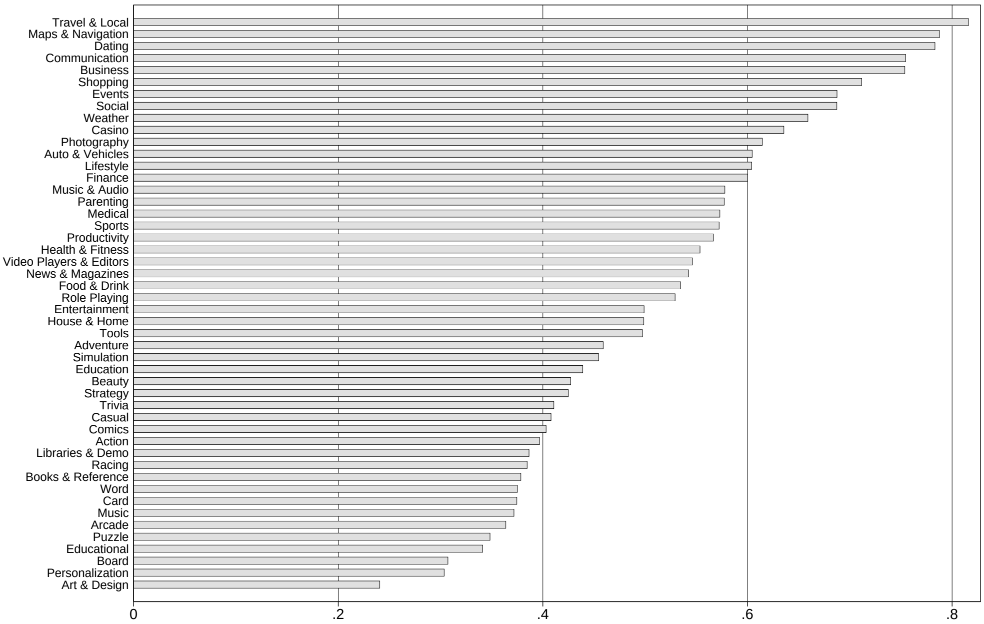


Figure 9: Average market concentration (HHI) by category

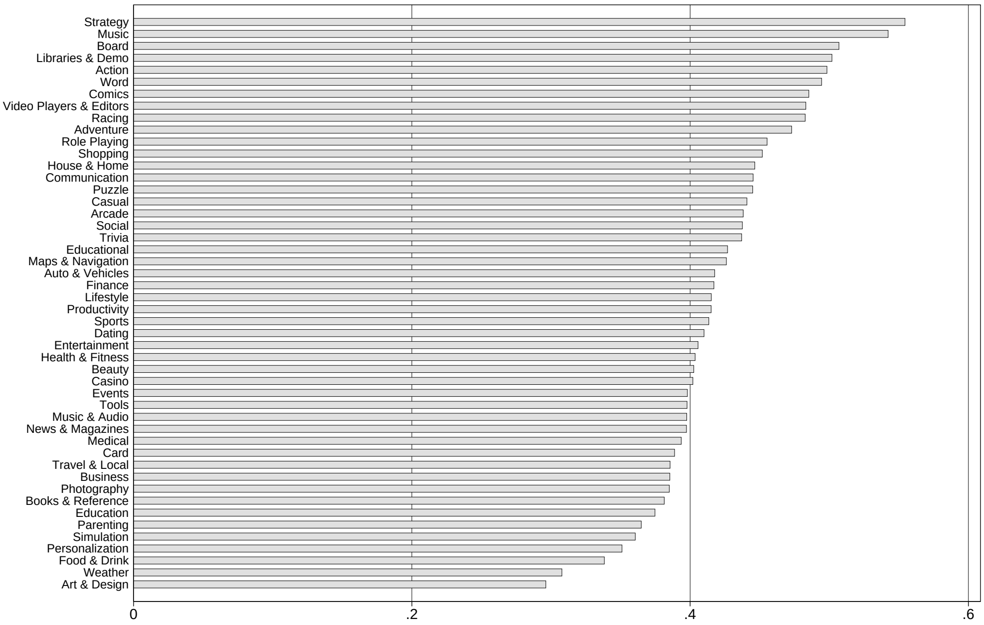


Table 3: Apps in categories before and after recategorization

July 2017			October 2017		
Category	N	%	Category	N	%
Action	19,600	1.63	Action	20,798	1.57
Adventure	12,254	1.02	Adventure	13,223	1.00
Arcade	47,072	3.90	Arcade	48,538	3.67
			Art & Design[a]	155	0.01
			Auto & Vehicles[b]	513	0.04
			Beauty[b]	221	0.02
Board	5,610	0.47	Board	5,828	0.44
Books & Reference	55,751	4.62	Books & Reference	60,639	4.59
Business	42,917	3.56	Business	67,176	5.08
Card	6,482	0.54	Card	6,735	0.51
Casino	4,831	0.40	Casino	5,168	0.39
Casual	50,093	4.16	Casual	51,893	3.93
Comics	2,620	0.22	Comics	2,746	0.21
Communication	27,733	2.30	Communication	29,060	2.20
			Dating[c]	193	0.01
Education	102,726	8.52	Education	107,633	8.14
Educational	17,249	1.43	Educational	18,031	1.36
Entertainment[d]	85,933	7.13	Entertainment	91,398	6.92
			Events[d]	127	0.01
Finance	27,775	2.30	Finance	28,686	2.17
			Food & Drink[b]	1,365	0.10
Health & Fitness	36,297	3.01	Health & Fitness	39,102	2.96
			House & Home[b]	223	0.02
Libraries & Demo	3,140	0.26	Libraries & Demo	3,132	0.24
Lifestyle[b]	87,984	7.30	Lifestyle	92,376	6.99
			Maps & Navigation	20,981	1.59
Media & Video	14,665	1.22			
Medical	14,539	1.21	Medical	17,682	1.34
Music	1,679	0.14	Music	1,738	0.13
Music & Audio	52,600	4.36	Music & Audio	55,508	4.20
News & Magazines	34,014	2.82	News & Magazines	35,430	2.68
			Parenting[b]	177	0.01
Personalization	83,068	6.89	Personalization	87,264	6.60
Photography	15,433	1.28	Photography	24,592	1.86
Productivity[a]	38,462	3.19	Productivity	40,155	3.04
Puzzle	46,925	3.89	Puzzle	52,894	4.00
Racing	8,433	0.70	Racing	10,535	0.80
Role Playing	1,846	0.15	Role Playing	4,323	0.33
Shopping	23,411	1.94	Shopping	25,497	1.93
Simulation	4,502	0.37	Simulation	14,913	1.13
Social[c]	26,795	2.22	Social	27,772	2.10
Sports	28,541	2.37	Sports	34,259	2.59
Strategy	4,929	0.41	Strategy	5,383	0.41
Tools	80,891	6.71	Tools	82,278	6.23
Transportation	20,559	1.71			
Travel & Local	49,578	4.11	Travel & Local	50,840	3.85
Trivia	8,150	0.68	Trivia	8,737	0.66
			Video Players & Editors	14,996	1.13
Weather	5,932	0.49	Weather	6,024	0.46
Word	4,411	0.37	Word	4,577	0.35
Total	1,205,430	100.00	Total	1,321,514	100.00

Notes: New categories are highlighted in bold. a, b, c, and d denote the categories in which apps were before the change (according to Google). In the regressions, these categories are called pre-move categories and are denoted by OLD.

Table 4: Alternative data collection measures

	$D_{DataCollection}$				$\#_{Category-SpecificDataCollection}$			
	CS1	CS2	Panel1	Panel2	CS1	CS2	Panel1	Panel2
HHI	0.050***	0.050***	0.004***	0.003***	0.077***	0.072***	0.002	0.002***
	(0.002)	(0.002)	(0.001)	(0.000)	(0.004)	(0.004)	(0.001)	(0.001)
Log. Market Share	0.002***	0.002***	0.001***	0.001***	0.006***	0.006***	0.001***	0.001***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Log. Ratings	-0.016***	-0.016***	-0.010***	-0.010***	-0.027***	-0.027***	-0.010***	-0.010***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
D_{Paid}	-0.098***	-0.098***	-0.014***	-0.014***	-0.176***	-0.176***	-0.030***	-0.030***
	(0.002)	(0.002)	(0.003)	(0.003)	(0.003)	(0.003)	(0.007)	(0.007)
$\#_{CleanPerms.}$	0.063***	0.063***	0.063***	0.063***	0.103***	0.103***	0.128***	0.128***
	(0.000)	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
$D_{InAppProduct}$	0.036***	0.036***	-0.084***	-0.084***	-0.037***	-0.037***	-0.113***	-0.113***
	(0.001)	(0.001)	(0.003)	(0.003)	(0.003)	(0.003)	(0.005)	(0.005)
Avg. Rating	-0.007***	-0.007***	-0.004***	-0.004***	-0.011***	-0.011***	-0.002***	-0.002***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.001)	(0.000)	(0.000)
Log. Length Description	-0.007***	-0.007***	0.007***	0.007***	-0.017***	-0.017***	0.014***	0.014***
	(0.000)	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)
$D_{IncludesAds}$	-0.080***	-0.080***	-0.026***	-0.026***	-0.092***	-0.092***	-0.027***	-0.027***
	(0.001)	(0.001)	(0.000)	(0.000)	(0.002)	(0.002)	(0.000)	(0.000)
Content Rating	0.009***	0.009***	-0.002***	-0.002***	0.017***	0.017***	-0.004***	-0.004***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)
$D_{PrivacyProfile}$	0.208***	0.208***	-0.009***	-0.009***	0.156***	0.156***	-0.017***	-0.017***
	(0.001)	(0.001)	(0.000)	(0.000)	(0.002)	(0.002)	(0.001)	(0.001)
$\#_{AppsbyDeveloper}$	-0.000***	-0.000***	0.000***	0.000***	0.000***	0.000***	0.000***	0.000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
App Age	0.000***	0.000***	0.000	0.000	0.000***	0.000***	0.000	0.000
	(0.000)	(0.000)	(.)	(.)	(0.000)	(0.000)	(.)	(.)
Constant	0.145***	0.145***	0.247***	0.247***	-0.014	-0.014	-0.239***	-0.239***
	(0.005)	(0.005)	(0.007)	(0.007)	(0.009)	(0.009)	(0.015)	(0.015)
Category	Yes	Yes	No	No	Yes	Yes	No	No
Wave	No	No	Yes	Yes	No	No	Yes	Yes
Mean $\#_{DataCollection}$	0.520	0.520	0.534	0.534	0.402	0.402	0.409	0.409
Observations	1,336,625	1,336,625	11,477,730	11,477,730	1,336,625	1,336,625	11,477,730	11,477,730
Num. of Groups			1,705,215	1,705,215			1,705,215	1,705,215
Adjusted R^2	0.36	0.36	0.14	0.14	0.29	0.29	0.16	0.16

Notes: The table shows the baseline estimations, when using alternative measures of data collection. The dependent variable in columns 1-4 is a dummy variable that takes the value of one if an app collects *any* privacy-sensitive permissions ($D_{DataCollection}$), and in columns 5-8, we use context-specific criteria based on the app categories to determine which permissions are privacy-sensitive and not necessarily functional ($Category - SpecificDataCollection$). The coefficients of interest are *HHI* and *Log. Market Share*. Heteroscedasticity-robust standard errors in parentheses: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 5: Alternative competition measures

	$\#_{DataCollection}$					
	CS1	Panel1	CS1	Panel1	CS1	Panel1
HHI 0.3-0.6	0.037***	0.001				
	(0.003)	(0.001)				
HHI 0.6-1	0.035***	0.001				
	(0.007)	(0.001)				
$\#_{NumberofCompetitors}/100$			0.001***	-0.002***		
			(0.000)	(0.000)		
Market Share of Top 4					0.053***	0.004***
					(0.005)	(0.001)
Log. Market Share	0.013***	0.001***	0.017***	-0.000	0.012***	0.001***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Log. Ratings	-0.079***	-0.019***	-0.082***	-0.017***	-0.078***	-0.019***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
D_{Paid}	-0.330***	0.028***	-0.329***	0.028***	-0.330***	0.027***
	(0.004)	(0.009)	(0.004)	(0.009)	(0.004)	(0.009)
$\#_{CleanPerms.}$	0.312***	0.267***	0.312***	0.266***	0.311***	0.266***
	(0.001)	(0.002)	(0.001)	(0.002)	(0.001)	(0.002)
$D_{InAppProduct}$	-0.118***	-0.025***	-0.119***	-0.023***	-0.117***	-0.026***
	(0.004)	(0.006)	(0.004)	(0.006)	(0.004)	(0.006)
Avg. Rating	-0.000	-0.008***	-0.001	-0.008***	-0.000	-0.008***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Log. Length Description	-0.035***	-0.017***	-0.035***	-0.017***	-0.035***	-0.017***
	(0.001)	(0.003)	(0.001)	(0.003)	(0.001)	(0.003)
$D_{IncludesAds}$	-0.321***	-0.046***	-0.322***	-0.045***	-0.324***	-0.047***
	(0.003)	(0.001)	(0.003)	(0.001)	(0.003)	(0.001)
Content Rating	0.013***	-0.006***	0.014***	-0.006***	0.013***	-0.006***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
$D_{PrivacyProfile}$	0.373***	-0.037***	0.372***	-0.035***	0.376***	-0.036***
	(0.003)	(0.001)	(0.003)	(0.001)	(0.003)	(0.001)
$\#_{AppsbyDeveloper}$	-0.000***	0.000***	-0.000***	0.000***	-0.000***	0.000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
App Age	0.000***	0.000	0.000***	0.000	0.000***	0.000
	(0.000)	(.)	(0.000)	(.)	(0.000)	(.)
Constant	-0.194***	0.264***	-0.146***	0.265***	-0.216***	0.263***
	(0.014)	(0.020)	(0.014)	(0.020)	(0.015)	(0.020)
Wave	No	Yes	No	Yes	No	Yes
Category	Yes	No	Yes	No	Yes	No
Mean $\#_{DataCollection}$	1.325	1.356	1.325	1.356	1.321	1.351
Observations	1,336,625	11,477,730	1,336,625	11,477,730	1,315,680	11,287,307
Num. of Groups		1,705,215		1,705,215		1,683,153
Adjusted R^2	0.52	0.28	0.52	0.28	0.52	0.28

Notes: The table shows the baseline estimations, when using alternative measures of competition. The dependent variable in these specifications is the number of privacy-sensitive permissions that an app collects *# DataCollection*. The coefficients of interest that quantify market concentration are significantly concentrated (*HHI = 0.3-0.6*) and highly concentrated (*HHI = 0.6-1.0*) markets (in columns 1-2), the number of competitors in the market ($\#_{NumberofCompetitors}$ in hundreds in columns 3-4) and the *Market Share of Top 4*. Heteroscedasticity-robust standard errors in parentheses: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 6: Explaining data sharing

	$D_{IntrusiveLibrary}$			
	CS1	CS2	CS1	CS2
HHI	0.012***	0.010***	0.005	0.005
	(0.004)	(0.004)	(0.004)	(0.004)
Log. Market Share			0.001***	0.001***
			(0.000)	(0.000)
Log. Ratings	0.011***	0.011***	0.010***	0.010***
	(0.000)	(0.000)	(0.000)	(0.000)
D_{Paid}	-0.118***	-0.118***	-0.118***	-0.118***
	(0.003)	(0.003)	(0.003)	(0.003)
$\#_{CleanPerms.}$	0.025***	0.025***	0.025***	0.025***
	(0.000)	(0.000)	(0.000)	(0.000)
$D_{InAppProduct}$	0.023***	0.023***	0.023***	0.023***
	(0.002)	(0.002)	(0.002)	(0.002)
Avg. Rating	0.003***	0.003***	0.003***	0.003***
	(0.001)	(0.001)	(0.001)	(0.001)
Log. Length Description	0.005***	0.005***	0.005***	0.005***
	(0.001)	(0.001)	(0.001)	(0.001)
$D_{IncludesAds}$	0.500***	0.500***	0.501***	0.501***
	(0.001)	(0.001)	(0.001)	(0.001)
Content Rating	-0.004***	-0.004***	-0.004***	-0.004***
	(0.001)	(0.001)	(0.001)	(0.001)
$D_{PrivacyProfile}$	0.025***	0.025***	0.026***	0.026***
	(0.002)	(0.002)	(0.002)	(0.002)
$\#_{AppsbyDeveloper}$	-0.000***	-0.000***	-0.000***	-0.000***
	(0.000)	(0.000)	(0.000)	(0.000)
App Age	-0.000***	-0.000***	-0.000***	-0.000***
	(0.000)	(0.000)	(0.000)	(0.000)
Constant	0.321***	0.321***	0.339***	0.340***
	(0.008)	(0.008)	(0.009)	(0.009)
Category	Yes	Yes	Yes	Yes
Mean $D_{IntrusiveLibrary}$	0.671	0.671	0.671	0.671
Observations	333,613	333,613	333,613	333,613
Adjusted R^2	0.45	0.45	0.45	0.45

Notes: The table analyzes the relationship between competition and data sharing as measured by the existence of intrusive software libraries within apps. The table shows cross-sectional estimates and the dependent variable is a dummy variable, which is equal to one if an app uses at least one intrusive software library ($D_{IntrusiveLibrary}$). The coefficients of interest are *HHI* and *Log. Market Share*. Heteroscedasticity-robust standard errors in parentheses: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 7: Baseline estimations (with balanced panel)

	$\#_{DataCollection}$							
	CS1	CS2	Panel1	Panel2	CS1	CS2	Panel1	Panel2
HHI	0.169***	0.154***	0.008***	0.008***	0.111***	0.109***	0.015***	0.015***
	(0.012)	(0.011)	(0.002)	(0.002)	(0.012)	(0.011)	(0.002)	(0.002)
Log. Market Share					0.012***	0.012***	0.003***	0.002***
					(0.001)	(0.001)	(0.000)	(0.000)
Log. Ratings	-0.075***	-0.075***	-0.017***	-0.017***	-0.085***	-0.085***	-0.020***	-0.020***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
D_{Paid}	-0.300***	-0.300***	0.061***	0.061***	-0.296***	-0.296***	0.061***	0.061***
	(0.009)	(0.009)	(0.017)	(0.017)	(0.009)	(0.009)	(0.017)	(0.017)
$\#_{CleanPerms.}$	0.324***	0.324***	0.261***	0.261***	0.324***	0.324***	0.261***	0.261***
	(0.001)	(0.001)	(0.004)	(0.004)	(0.001)	(0.001)	(0.004)	(0.004)
$D_{InAppProduct}$	-0.130***	-0.130***	-0.044***	-0.044***	-0.128***	-0.128***	-0.044***	-0.044***
	(0.006)	(0.006)	(0.010)	(0.010)	(0.006)	(0.006)	(0.010)	(0.010)
Avg. Rating	0.001	0.001	-0.009***	-0.009***	-0.001	-0.001	-0.009***	-0.009***
	(0.002)	(0.002)	(0.001)	(0.001)	(0.002)	(0.002)	(0.001)	(0.001)
Log. Length Description	-0.031***	-0.031***	-0.026***	-0.026***	-0.030***	-0.030***	-0.026***	-0.026***
	(0.002)	(0.002)	(0.005)	(0.005)	(0.002)	(0.002)	(0.005)	(0.005)
$D_{IncludesAds}$	-0.292***	-0.292***	-0.055***	-0.055***	-0.289***	-0.289***	-0.055***	-0.055***
	(0.004)	(0.004)	(0.001)	(0.001)	(0.004)	(0.004)	(0.001)	(0.001)
Content Rating	0.020***	0.020***	-0.007***	-0.007***	0.019***	0.019***	-0.007***	-0.007***
	(0.002)	(0.002)	(0.001)	(0.001)	(0.002)	(0.002)	(0.001)	(0.001)
$D_{PrivacyProfile}$	0.317***	0.317***	-0.043***	-0.043***	0.318***	0.318***	-0.043***	-0.043***
	(0.005)	(0.005)	(0.002)	(0.002)	(0.005)	(0.005)	(0.002)	(0.002)
$\#_{AppsbyDeveloper}$	-0.000***	-0.000***	0.000***	0.000***	-0.000***	-0.000***	0.000***	0.000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
App Age	0.000***	0.000***	0.000	0.000	0.000***	0.000***	0.000	0.000
	(0.000)	(0.000)	(.)	(.)	(0.000)	(0.000)	(.)	(.)
Constant	-0.422***	-0.422***	0.382***	0.382***	-0.258***	-0.259***	0.409***	0.407***
	(0.022)	(0.022)	(0.029)	(0.029)	(0.024)	(0.024)	(0.030)	(0.030)
Category	Yes	Yes	No	No	Yes	Yes	No	No
Wave	No	No	Yes	Yes	No	No	Yes	Yes
Mean $\#_{DataCollection}$	1.428	1.428	1.448	1.448	1.428	1.428	1.448	1.448
Observations	449,093	449,093	4,245,348	4,245,348	449,093	449,093	4,245,348	4,245,348
Num. of Groups			573,817	573,817			573,817	573,817
Adjusted R^2	0.52	0.52	0.28	0.28	0.52	0.52	0.28	0.28

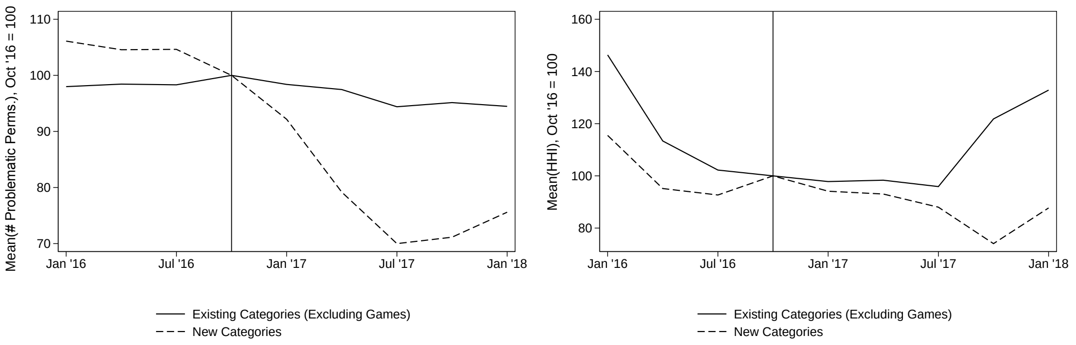
Notes: The table shows the baseline estimations, when restricting to a balanced panel. Cross-sectional and panel fixed effects regressions are given, while the dependent variable is the number of privacy-sensitive permissions that an app collects *# DataCollection*. The coefficients of interest are *HHI* and *Log. Market Share*. Heteroscedasticity-robust standard errors in parentheses: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 8: Categories defining the market

	Total Category				Top 100 of Category			
	CS1	CS2	Panel1	Panel2	CS1	CS2	Panel1	Panel2
HHI	3.152***	2.469***	-0.004	-0.009***	2.049***	1.750***	0.002	-0.000
	(0.040)	(0.030)	(0.003)	(0.003)	(0.422)	(0.360)	(0.045)	(0.041)
Log. Market Share	0.088***	0.084***	-0.004***	-0.004***	0.078***	0.074***	-0.003	-0.003
	(0.001)	(0.001)	(0.000)	(0.000)	(0.022)	(0.021)	(0.008)	(0.008)
Log. Ratings	-0.166***	-0.167***	-0.013***	-0.013***	-0.048***	-0.049***	0.015	0.015
	(0.001)	(0.001)	(0.001)	(0.001)	(0.016)	(0.016)	(0.026)	(0.026)
D_{Paid}	-0.428***	-0.428***	0.028***	0.028***	0.167	0.169	0.065	0.064
	(0.004)	(0.004)	(0.009)	(0.009)	(0.373)	(0.373)	(0.055)	(0.055)
$\#_{CleanPerms.}$	0.319***	0.319***	0.267***	0.267***	0.286***	0.286***	0.252***	0.252***
	(0.001)	(0.001)	(0.002)	(0.002)	(0.023)	(0.023)	(0.010)	(0.010)
$D_{InAppProduct}$	-0.162***	-0.159***	-0.025***	-0.025***	-0.259***	-0.260***	0.102**	0.102**
	(0.004)	(0.004)	(0.006)	(0.006)	(0.050)	(0.050)	(0.051)	(0.051)
Avg. Rating	-0.005***	-0.005***	-0.008***	-0.008***	-0.088	-0.087	-0.111	-0.112
	(0.001)	(0.001)	(0.001)	(0.001)	(0.054)	(0.054)	(0.089)	(0.089)
Log. Length Description	-0.033***	-0.033***	-0.017***	-0.017***	0.016	0.016	0.022	0.022
	(0.001)	(0.001)	(0.003)	(0.003)	(0.034)	(0.034)	(0.028)	(0.028)
$D_{IncludesAds}$	-0.404***	-0.402***	-0.046***	-0.046***	-0.622***	-0.623***	-0.088***	-0.088***
	(0.002)	(0.002)	(0.001)	(0.001)	(0.062)	(0.062)	(0.017)	(0.017)
Content Rating	0.005***	0.005***	-0.007***	-0.007***	-0.005	-0.006	0.028	0.028
	(0.001)	(0.001)	(0.001)	(0.001)	(0.021)	(0.021)	(0.017)	(0.017)
$D_{PrivacyProfile}$	0.382***	0.382***	-0.037***	-0.037***	0.536***	0.536***	0.015	0.015
	(0.004)	(0.004)	(0.001)	(0.001)	(0.079)	(0.079)	(0.018)	(0.018)
$\#_{AppsbyDeveloper}$	-0.000***	-0.000***	0.000***	0.000***	-0.000	-0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
App Age	0.000***	0.000***	0.000	0.000	0.000***	0.000***	0.000	0.000
	(0.000)	(0.000)	(.)	(.)	(0.000)	(0.000)	(.)	(.)
Constant	1.265***	1.259***	0.174***	0.176***	0.569	0.556	0.143	0.144
	(0.028)	(0.028)	(0.021)	(0.021)	(0.387)	(0.384)	(0.446)	(0.445)
Wave	No	No	Yes	Yes	No	No	Yes	Yes
Mean $\#_{DataCollection}$	1.325	1.325	1.356	1.356	2.048	2.048	2.066	2.066
Observations	1,336,625	1,336,625	11,477,730	11,477,730	4,800	4,800	40,509	40,509
Num. of Groups			1,705,215	1,705,215			15,306	15,306
Adjusted R^2	0.51	0.51	0.28	0.28	0.61	0.61	0.27	0.27

Notes: The table analyzes the relationship between market concentration and data collection with the market being defined by the category or by the 100 most popular apps in the category. Cross-sectional and panel fixed effects regressions are given, while the dependent variable is the number of privacy-sensitive permissions that an app collects *# DataCollection*. The coefficients of interest are *HHI* and *Log. Market Share*. Heteroscedasticity-robust standard errors in parentheses: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Figure 10: Development of new and existing categories



(a) Number of Problematic Permissions

(b) HHI

Notes: Based on the whole observation period, the figures show the development of key variables around the recategorization comparing new and existing categories: Panel (a) shows the development for the number of problematic permissions requested on average, while panel (b) shows the average HHI. In both cases, the value at the time of recategorization is set as a benchmark.

Table 9: Summary statistics comparing new and existing categories

	Existing Categories (Excluding Games)					New Categories				
	Mean	P50	Min	Max	Count	Mean	P50	Min	Max	Count
$D_{DataCollection}$	0.58	1.00	0	1	1,008,325	0.69	1.00	0	1	5,639
$\#_{DataCollection}$	1.57	1.00	0	15	1,008,325	2.01	2.00	0	12	5,639
HHI	0.12	0.06	0	1	1,008,325	0.24	0.15	0	1	5,639
HHI	0.13	0.07	0	1	1,008,325	0.26	0.16	0	1	5,639
Market Share	0.02	0.00	0	1	1,008,325	0.07	0.00	0	1	5,639
Market Share	0.02	0.00	0	1	1,008,325	0.07	0.00	0	1	5,639
D_{Paid}	0.08	0.00	0	1	1,008,325	0.04	0.00	0	1	5,639
$\#_{CleanPerms.}$	4.72	4.00	0	169	1,008,325	5.90	6.00	0	27	5,639
$D_{InAppProduct}$	0.06	0.00	0	1	1,008,325	0.10	0.00	0	1	5,639
Avg. Rating	4.08	4.20	1	5	1,008,325	4.20	4.30	1	5	5,639
Length Description	1,065.59	740.00	1	11,588	1,008,325	1,457.37	1,021.00	3	11,034	5,639
Content Rating	8.25	8.00	3	10	1,008,325	7.69	8.00	3	10	5,639
$D_{PrivacyProfile}$	0.18	0.00	0	1	1,008,325	0.41	0.00	0	1	5,639
$\#_{AppsbyDeveloper}$	71.51	8.00	1	2,490	1,008,325	91.84	7.00	1	1,270	5,639
App Age	1,229.36	1,119.00	-15	3,334	1,008,325	1,037.66	930.00	43	2,925	5,639

Notes: The table compares summary statistics between new and existing categories at the time of recategorization, the former comprising all apps that chose to switch. The variable $D_{IncludesAds}$ is excluded as back then it was not mandatory to indicate, whether ads are included.

Table 10: Recategorization (pooled OLS)

	$\#_{DataCollection}$				
	ALL	OLD	NOT	OLD-BP	OLD-CL
HHI	0.041***	0.083***	0.020***	0.112***	0.014**
	(0.003)	(0.006)	(0.004)	(0.010)	(0.007)
Log. Market Share	0.015***	0.014***	0.014***	0.010***	0.019***
	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)
Recat x t_{Recat}-2	0.034	0.050*	0.031	0.080*	0.057**
	(0.026)	(0.026)	(0.026)	(0.044)	(0.026)
Recat x t_{Recat}-1	-0.012	-0.012	-0.016	-0.005	-0.005
	(0.024)	(0.024)	(0.024)	(0.038)	(0.025)
Recat x t_{Recat}	-0.015	-0.041**	-0.005	-0.052	-0.029
	(0.020)	(0.020)	(0.020)	(0.035)	(0.020)
Recat x t_{Recat}+1	-0.015	-0.036**	-0.007	-0.026	-0.028
	(0.018)	(0.018)	(0.018)	(0.035)	(0.018)
Recat x t_{Recat}+2	-0.041***	-0.076***	-0.031**	-0.082**	-0.075***
	(0.016)	(0.016)	(0.016)	(0.036)	(0.016)
Recat x t_{Recat}+3	-0.040***	-0.057***	-0.037**	-0.106***	-0.056***
	(0.015)	(0.016)	(0.016)	(0.039)	(0.016)
Recat x t_{Recat}+4	-0.033**	-0.049***	-0.026*	-0.073**	-0.053***
	(0.016)	(0.016)	(0.016)	(0.037)	(0.016)
Recat	-0.062***	-0.014	-0.096***	-0.049	-0.011
	(0.015)	(0.018)	(0.018)	(0.033)	(0.018)
t_{Recat}-2	-0.023***	-0.038***	-0.017***	-0.048***	-0.034***
	(0.002)	(0.004)	(0.002)	(0.006)	(0.005)
t_{Recat}-1	-0.020***	-0.036***	-0.012***	-0.042***	-0.032***
	(0.002)	(0.004)	(0.002)	(0.006)	(0.004)
t_{Recat}	-0.073***	-0.058***	-0.090***	-0.049***	-0.068***
	(0.002)	(0.005)	(0.003)	(0.007)	(0.006)
t_{Recat}+1	-0.070***	-0.070***	-0.073***	-0.051***	-0.065***
	(0.002)	(0.004)	(0.002)	(0.006)	(0.004)
t_{Recat}+2	0.018***	0.035***	0.012***	0.037***	0.026***
	(0.002)	(0.003)	(0.002)	(0.006)	(0.004)
t_{Recat}+3	0.007***	0.025***	-0.000	0.035***	0.015***
	(0.002)	(0.003)	(0.002)	(0.006)	(0.004)
t_{Recat}+4	0.004**	0.028***	-0.004**	0.034***	0.018***
	(0.002)	(0.003)	(0.002)	(0.006)	(0.004)
Log. Ratings	-0.086***	-0.112***	-0.076***	-0.113***	-0.117***
	(0.000)	(0.001)	(0.000)	(0.001)	(0.001)
D_{Paid}	-0.267***	-0.342***	-0.242***	-0.346***	-0.324***
	(0.002)	(0.004)	(0.002)	(0.008)	(0.005)
$\#_{CleanPerms.}$	0.328***	0.355***	0.319***	0.355***	0.352***
	(0.000)	(0.001)	(0.001)	(0.002)	(0.001)
$D_{InAppProduct}$	-0.146***	-0.079***	-0.161***	-0.076***	-0.139***
	(0.002)	(0.004)	(0.002)	(0.007)	(0.005)
Avg. Rating	0.008***	0.014***	0.003***	0.019***	0.014***
	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)
Log. Length Description	-0.025***	-0.006***	-0.031***	-0.009***	-0.000
	(0.000)	(0.001)	(0.001)	(0.002)	(0.001)
$D_{IncludesAds}$	-0.346***	-0.386***	-0.325***	-0.320***	-0.354***
	(0.001)	(0.003)	(0.001)	(0.005)	(0.003)
Content Rating	0.014***	0.006***	0.021***	0.006***	0.011***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
$D_{PrivacyProfile}$	0.288***	0.341***	0.267***	0.331***	0.380***
	(0.001)	(0.003)	(0.001)	(0.005)	(0.004)
$\#_{AppsbyDeveloper}$	0.000***	0.000***	0.000***	-0.000***	0.000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
App Age	0.000***	0.000***	0.000***	0.000***	0.000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Constant	0.193	1.211***	0.247	-1.116***	1.205***
	(0.229)	(0.025)	(0.229)	(0.059)	(0.026)
Category	Yes	Yes	Yes	Yes	Yes
Mean $\#_{DataCollection}$	1.524	1.645	1.484	1.770	1.612
Observations	8,886,685	2,274,152	6,683,836	880,884	1,514,417
Num. of Groups					
Adjusted R^2	0.52	0.50	0.53	0.48	0.50

Notes: The table shows pooled OLS regressions, while the dependent variable is the number of privacy-sensitive permissions that an app requests *# DataCollection*. The coefficients of interest are the interactions between *Recat*, a dummy equal to one for apps that moved into new categories, and time dummies denoted by t_{Recat}. Column titles denote the control group and sample (ALL = all apps, OLD = pre-move categories, NOT = non-move categories, BP = balanced panel, CL = cluster with movers omitted). Heteroscedasticity-robust standard errors in parentheses: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 11: Recategorization (app fixed effects)

	$\#_{DataCollection}$				
	ALL	OLD	NOT	OLD-BP	OLD-CL
HHI	0.027***	0.028***	0.027***	0.033***	0.022***
	(0.002)	(0.003)	(0.002)	(0.005)	(0.004)
Log. Market Share	0.007***	0.006***	0.007***	0.007***	0.004***
	(0.000)	(0.001)	(0.000)	(0.001)	(0.001)
Recat x t_{Recat}-2	0.044***	0.055***	0.045***	0.121***	0.057***
	(0.011)	(0.012)	(0.011)	(0.022)	(0.011)
Recat x t_{Recat}-1	-0.007	-0.001	-0.006	0.019	0.000
	(0.009)	(0.009)	(0.009)	(0.017)	(0.009)
Recat x t_{Recat}	-0.021***	-0.018**	-0.022***	-0.021	-0.016**
	(0.008)	(0.008)	(0.008)	(0.015)	(0.008)
Recat x t_{Recat}+1	-0.022***	-0.029***	-0.022***	-0.065***	-0.028***
	(0.007)	(0.007)	(0.007)	(0.014)	(0.007)
Recat x t_{Recat}+2	-0.014**	-0.019***	-0.012**	-0.020	-0.018***
	(0.006)	(0.006)	(0.006)	(0.013)	(0.006)
Recat x t_{Recat}+3	-0.009*	-0.014***	-0.008*	-0.027**	-0.013**
	(0.005)	(0.005)	(0.005)	(0.012)	(0.005)
Recat x t_{Recat}+4	-0.020***	-0.031***	-0.018***	-0.100***	-0.030***
	(0.004)	(0.005)	(0.004)	(0.013)	(0.005)
t_{Recat}-2	0.012***	0.012***	0.011***	0.012***	0.010***
	(0.000)	(0.001)	(0.000)	(0.001)	(0.001)
t_{Recat}-1	0.008***	0.012***	0.007***	0.013***	0.010***
	(0.000)	(0.001)	(0.000)	(0.001)	(0.001)
t_{Recat}	0.025***	0.039***	0.018***	0.037***	0.050***
	(0.003)	(0.007)	(0.004)	(0.009)	(0.009)
t_{Recat}+1	-0.001	0.012***	-0.006***	0.016***	0.017***
	(0.001)	(0.002)	(0.001)	(0.003)	(0.003)
t_{Recat}+2	-0.009***	-0.005***	-0.010***	-0.001	-0.004***
	(0.000)	(0.001)	(0.001)	(0.002)	(0.001)
t_{Recat}+3	-0.004***	0.000	-0.006***	0.005***	-0.001
	(0.000)	(0.001)	(0.000)	(0.001)	(0.001)
t_{Recat}+4	-0.010***	-0.002**	-0.013***	-0.000	-0.001
	(0.000)	(0.001)	(0.000)	(0.001)	(0.001)
Log. Ratings	-0.028***	-0.033***	-0.026***	-0.038***	-0.029***
	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)
D_{Paid}	0.043***	0.006	0.053***	-0.035	0.001
	(0.011)	(0.021)	(0.012)	(0.040)	(0.024)
$\#_{CleanPerms.}$	0.271***	0.272***	0.271***	0.259***	0.273***
	(0.002)	(0.007)	(0.003)	(0.015)	(0.010)
$D_{InAppProduct}$	-0.086***	0.014	-0.118***	-0.032	-0.070***
	(0.008)	(0.019)	(0.009)	(0.031)	(0.023)
Avg. Rating	-0.005***	-0.009***	-0.003***	-0.012***	-0.009***
	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)
Log. Length Description	-0.017***	-0.027***	-0.013***	-0.050***	-0.030***
	(0.003)	(0.006)	(0.004)	(0.010)	(0.008)
$D_{IncludesAds}$	-0.063***	-0.043***	-0.070***	-0.045***	-0.040***
	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)
Content Rating	-0.007***	-0.010***	-0.005***	-0.008***	-0.014***
	(0.001)	(0.002)	(0.001)	(0.003)	(0.003)
$D_{PrivacyProfile}$	-0.036***	-0.044***	-0.034***	-0.051***	-0.042***
	(0.002)	(0.003)	(0.002)	(0.005)	(0.004)
$\#_{AppsbyDeveloper}$	0.000***	0.000*	0.000***	0.000**	0.000*
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Constant	0.435**	0.586***	0.027	0.971***	0.615***
	(0.190)	(0.065)	(0.343)	(0.124)	(0.083)
Category	Yes	Yes	Yes	Yes	Yes
Mean $\#_{DataCollection}$	1.524	1.645	1.484	1.770	1.612
Observations	8,886,685	2,274,152	6,683,836	880,884	1,514,417
Num. of Groups	1,312,919	341,019	991,495	119,012	228,270
Adjusted R^2	0.29	0.26	0.29	0.25	0.26

Notes: The table shows panel regressions with app fixed effects, while the dependent variable is the number of privacy-sensitive permissions that an app requests *# DataCollection*. The coefficients of interest are the interactions between *Recat*, a dummy equal to one for apps that moved into new categories, and time dummies denoted by t_{Recat}. Column titles denote the control group and sample (ALL = all apps, OLD = pre-move categories, NOT = non-move categories, BP = balanced panel, CL = cluster with movers omitted). Heteroscedasticity-robust standard errors in parentheses: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 12: Contrasting free and paid apps

	Free		Paid		Free		Paid	
	CS1	Panel1	CS1	Panel1	CS2	Panel2	CS2	Panel2
HHI	0.089***	0.009***	0.020	0.005*	0.088***	0.008***	0.030	0.005
	(0.007)	(0.001)	(0.024)	(0.003)	(0.007)	(0.001)	(0.022)	(0.003)
Log. Market Share	0.012***	0.002***	0.013***	0.001	0.012***	0.002***	0.013***	0.001
	(0.000)	(0.000)	(0.001)	(0.001)	(0.000)	(0.000)	(0.001)	(0.000)
Log. Ratings	-0.079***	-0.019***	-0.049***	-0.020***	-0.079***	-0.019***	-0.049***	-0.020***
	(0.001)	(0.001)	(0.003)	(0.004)	(0.001)	(0.001)	(0.003)	(0.004)
$\#_{CleanPerms.}$	0.317***	0.269***	0.224***	0.179***	0.317***	0.269***	0.224***	0.179***
	(0.001)	(0.002)	(0.003)	(0.018)	(0.001)	(0.002)	(0.003)	(0.018)
$D_{InAppProduct}$	-0.125***	-0.033***	0.018	0.227***	-0.125***	-0.033***	0.018	0.227***
	(0.004)	(0.006)	(0.019)	(0.033)	(0.004)	(0.006)	(0.019)	(0.033)
Avg. Rating	0.002	-0.008***	-0.042***	-0.001	0.002	-0.008***	-0.042***	-0.001
	(0.002)	(0.001)	(0.003)	(0.002)	(0.002)	(0.001)	(0.003)	(0.002)
Log. Length Description	-0.031***	-0.018***	-0.056***	-0.006	-0.031***	-0.018***	-0.056***	-0.006
	(0.001)	(0.003)	(0.004)	(0.011)	(0.001)	(0.003)	(0.004)	(0.011)
$D_{IncludesAds}$	-0.322***	-0.046***	0.106***	0.003	-0.322***	-0.046***	0.106***	0.003
	(0.003)	(0.001)	(0.015)	(0.003)	(0.003)	(0.001)	(0.015)	(0.003)
Content Rating	0.008***	-0.006***	0.051***	-0.002	0.008***	-0.006***	0.051***	-0.002
	(0.001)	(0.001)	(0.005)	(0.001)	(0.001)	(0.001)	(0.005)	(0.001)
$D_{PrivacyProfile}$	0.377***	-0.037***	0.261***	-0.031***	0.377***	-0.037***	0.261***	-0.031***
	(0.004)	(0.001)	(0.009)	(0.004)	(0.004)	(0.001)	(0.009)	(0.004)
$\#_{AppsbyDeveloper}$	-0.000***	0.000***	-0.000***	0.000***	-0.000***	0.000***	-0.000***	0.000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
App Age	0.000***	0.000	0.000***	0.000	0.000***	0.000	0.000***	0.000
	(0.000)	(.)	(0.000)	(.)	(0.000)	(.)	(0.000)	(.)
Constant	-0.260***	0.281***	-0.002	0.192**	-0.259***	0.280***	-0.008	0.191**
	(0.015)	(0.020)	(0.049)	(0.076)	(0.015)	(0.020)	(0.048)	(0.076)
Wave	No	Yes	No	Yes	No	Yes	No	Yes
Category	Yes	No	Yes	No	Yes	No	Yes	No
Mean $\#_{DataCollection}$	1.370	1.405	0.688	0.695	1.370	1.405	0.688	0.695
Observations	1,249,591	10,688,842	87,034	788,888	1,249,591	10,688,842	87,034	788,888
Num. of Groups		1,597,521		111,463		1,597,521		111,463
Adjusted R^2	0.53	0.28	0.42	0.19	0.53	0.28	0.42	0.19

Notes: The table shows the baseline estimations for different subgroups of apps. Columns 1, 2, 5, and 6 analyze free apps, while columns 3, 4, 7, and 8 show the relationship for paid apps. It shows only panel regressions and the dependent variable is the number of privacy-sensitive permissions that an app collects *# DataCollection*. The coefficients of interest are *HHI* and *Log. Market Share*. Heteroscedasticity-robust standard errors in parentheses: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 13: Restricting to large and important markets

	More than 10 Apps				More than 100,000 Installations			
	CS1	CS2	Panel1	Panel2	CS1	CS2	Panel1	Panel2
HHI	0.143***	0.132***	0.023***	0.021***	0.085***	0.082***	0.014***	0.013***
	(0.012)	(0.010)	(0.002)	(0.002)	(0.009)	(0.008)	(0.002)	(0.001)
Log. Market Share	0.024***	0.023***	0.004***	0.003***	0.012***	0.011***	0.003***	0.003***
	(0.001)	(0.001)	(0.000)	(0.000)	(0.001)	(0.001)	(0.000)	(0.000)
Log. Ratings	-0.085***	-0.085***	-0.021***	-0.021***	-0.074***	-0.074***	-0.021***	-0.020***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
D_{Paid}	-0.333***	-0.333***	0.047***	0.047***	-0.338***	-0.338***	0.029***	0.029***
	(0.005)	(0.005)	(0.011)	(0.011)	(0.005)	(0.005)	(0.010)	(0.010)
$\#_{CleanPerms.}$	0.304***	0.304***	0.256***	0.256***	0.302***	0.302***	0.262***	0.262***
	(0.001)	(0.001)	(0.003)	(0.003)	(0.001)	(0.001)	(0.002)	(0.002)
$D_{InAppProduct}$	-0.117***	-0.117***	-0.036***	-0.036***	-0.106***	-0.106***	-0.027***	-0.027***
	(0.005)	(0.005)	(0.008)	(0.008)	(0.004)	(0.004)	(0.006)	(0.006)
Avg. Rating	0.003	0.003	-0.007***	-0.007***	-0.002	-0.002	-0.008***	-0.008***
	(0.002)	(0.002)	(0.001)	(0.001)	(0.002)	(0.002)	(0.001)	(0.001)
Log. Length Description	-0.035***	-0.035***	-0.006*	-0.006*	-0.037***	-0.037***	-0.018***	-0.018***
	(0.001)	(0.001)	(0.004)	(0.004)	(0.001)	(0.001)	(0.003)	(0.003)
$D_{IncludesAds}$	-0.370***	-0.370***	-0.057***	-0.057***	-0.360***	-0.360***	-0.050***	-0.050***
	(0.003)	(0.003)	(0.001)	(0.001)	(0.003)	(0.003)	(0.001)	(0.001)
Content Rating	0.009***	0.009***	-0.011***	-0.011***	0.012***	0.012***	-0.008***	-0.008***
	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
$D_{PrivacyProfile}$	0.387***	0.387***	-0.039***	-0.039***	0.399***	0.399***	-0.038***	-0.038***
	(0.004)	(0.004)	(0.002)	(0.002)	(0.004)	(0.004)	(0.001)	(0.001)
$\#_{AppsbyDeveloper}$	-0.000***	-0.000***	0.000***	0.000***	-0.000***	-0.000***	0.000***	0.000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
App Age	0.000***	0.000***	0.000	0.000	0.000***	0.000***	0.000	0.000
	(0.000)	(0.000)	(.)	(.)	(0.000)	(0.000)	(.)	(.)
Constant	0.030	0.030	0.271***	0.268***	-0.122***	-0.120***	0.277***	0.274***
	(0.020)	(0.020)	(0.025)	(0.025)	(0.016)	(0.016)	(0.021)	(0.021)
Category	Yes	Yes	No	No	Yes	Yes	No	No
Wave	No	No	Yes	Yes	No	No	Yes	Yes
Mean $\#_{DataCollection}$	1.287	1.287	1.314	1.314	1.263	1.263	1.291	1.291
Observations	826,982	826,982	7,047,188	7,047,188	1,129,623	1,129,623	9,651,078	9,651,078
Num. of Groups			1,037,607	1,037,607			1,433,422	1,433,422
Adjusted R^2	0.53	0.53	0.26	0.26	0.52	0.52	0.27	0.27

Notes: The table shows the baseline estimations, when restricting to specific markets. Columns 1-4 include markets that consist of at least ten apps throughout the observation period, while columns 5-8 only consider those that have a minimum of 100,000 total installations at the beginning. Cross-sectional and panel regressions are given, while the dependent variable is the number of privacy-sensitive permissions that an app collects *# DataCollection*. The coefficients of interest are *HHI* and *Log. Market Share*. Heteroscedasticity-robust standard errors in parentheses: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 14: Varying the functional form of market shares

	$\#_{DataCollection}$							
	CS1	CS2	Panel1	Panel2	CS1	CS2	Panel1	Panel2
HHI	0.109***	0.098***	0.004***	0.004***	0.109***	0.100***	0.004***	0.004***
	(0.007)	(0.007)	(0.001)	(0.001)	(0.007)	(0.007)	(0.001)	(0.001)
Market Share	0.158***	0.162***	0.007	0.006	0.549***	0.551***	0.033***	0.030**
	(0.013)	(0.013)	(0.005)	(0.005)	(0.035)	(0.035)	(0.012)	(0.012)
Market Share Squared					-0.581***	-0.568***	-0.034**	-0.031**
					(0.049)	(0.048)	(0.015)	(0.014)
Log. Ratings	-0.070***	-0.070***	-0.018***	-0.018***	-0.070***	-0.071***	-0.018***	-0.018***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
D_{Paid}	-0.334***	-0.334***	0.028***	0.028***	-0.333***	-0.333***	0.028***	0.028***
	(0.004)	(0.004)	(0.009)	(0.009)	(0.004)	(0.004)	(0.009)	(0.009)
$\#_{CleanPerms.}$	0.312***	0.312***	0.267***	0.267***	0.312***	0.312***	0.267***	0.267***
	(0.001)	(0.001)	(0.002)	(0.002)	(0.001)	(0.001)	(0.002)	(0.002)
$D_{InAppProduct}$	-0.120***	-0.120***	-0.025***	-0.025***	-0.119***	-0.119***	-0.025***	-0.025***
	(0.004)	(0.004)	(0.006)	(0.006)	(0.004)	(0.004)	(0.006)	(0.006)
Avg. Rating	0.001	0.001	-0.008***	-0.008***	0.001	0.001	-0.008***	-0.008***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Log. Length Description	-0.035***	-0.035***	-0.017***	-0.017***	-0.035***	-0.035***	-0.017***	-0.017***
	(0.001)	(0.001)	(0.003)	(0.003)	(0.001)	(0.001)	(0.003)	(0.003)
$D_{IncludesAds}$	-0.324***	-0.324***	-0.046***	-0.046***	-0.323***	-0.323***	-0.046***	-0.046***
	(0.003)	(0.003)	(0.001)	(0.001)	(0.003)	(0.003)	(0.001)	(0.001)
Content Rating	0.015***	0.015***	-0.006***	-0.006***	0.014***	0.014***	-0.006***	-0.006***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
$D_{PrivacyProfile}$	0.372***	0.372***	-0.037***	-0.037***	0.372***	0.372***	-0.037***	-0.037***
	(0.003)	(0.003)	(0.001)	(0.001)	(0.003)	(0.003)	(0.001)	(0.001)
$\#_{AppsbyDeveloper}$	-0.000***	-0.000***	0.000***	0.000***	-0.000***	-0.000***	0.000***	0.000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
App Age	0.000***	0.000***	0.000	0.000	0.000***	0.000***	0.000	0.000
	(0.000)	(0.000)	(.)	(.)	(0.000)	(0.000)	(.)	(.)
Constant	-0.373***	-0.373***	0.250***	0.250***	-0.373***	-0.372***	0.250***	0.250***
	(0.013)	(0.013)	(0.020)	(0.020)	(0.013)	(0.013)	(0.020)	(0.020)
Category	Yes	Yes	No	No	Yes	Yes	No	No
Wave	No	No	Yes	Yes	No	No	Yes	Yes
Mean $\#_{DataCollection}$	1.325	1.325	1.356	1.356	1.325	1.325	1.356	1.356
Observations	1,336,625	1,336,625	1,1477,730	11,477,730	1,336,625	1,336,625	11,477,730	11,477,730
Num. of Groups			1,705,215	1,705,215			1,705,215	1,705,215
Adjusted R^2	0.52	0.52	0.28	0.28	0.52	0.52	0.28	0.28

Notes: The table shows the baseline estimations, when varying the functional form of markets. Columns 1-4 include the linear market share, while columns 5-8 additionally control for the squared market share. Cross-sectional and panel regressions are given, while the dependent variable is the number of privacy-sensitive permissions that an app collects *# DataCollection*. The coefficients of interest are *HHI*, *Market Share* and *Market Share Squared*. Heteroscedasticity-robust standard errors in parentheses: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

A.3 Exogenous variation in intensity of network effects

A.3.1 Instrumental variables based on network effects

To identify the effect of market power on data collection causally, we exploit variation in the intensity of a submarket's network effects (NE). Some services that apps provide can reach their full service quality independently of whether others use the same app or not (e.g., a wallpaper). These apps have no network effects. Other apps, like messengers or dating apps, are only useful if a user's relevant peers also use the same app. Such apps have very strong network effects. We can also find apps with intermediate network effects. These are apps that can provide basic functionality and utility independent of who else uses it, but have additional features that require the app's adoption by fellow users. An example could be a running app that lets you share some personal statistics with one's peers who also use the same app.

We employ several ways to measure the extent to which network effects are present in the respective market. The first approach exploits app characteristics to infer the extent through a badge on the app's Google Play Store page highlighting that users interact with others within the app. Second, we look at the content of the app's description, whether keywords appear that suggest the need for other users and thus the presence of network effects such as 'multiplayer,' 'friends,' and 'share.' Third, we randomly present users information from the Play Store on apps from each category and ask, whether it is beneficial if a certain app is also used by others or their friends.[29]

In this identification approach we exploit the fact that the strength of a submarket's network effects depends heavily on the nature of the service that is being provided, but not on the strategy of the players in the market. The strength of a service's network effects thus 'programs' the submarket's concentration in equilibrium and is also a good predictor of the observed market concentration, as it converges to the equilibrium state. While this phenomenon creates a research opportunity in and of itself, the strength of the network effects can also serve as a instrumental variable that exogenously affects concentration.

[29] So far, this has been only done with a few research assistants, but we plan to enlarge the sample by asking individuals at Amazon Mechanical Turk.

A.3.2 Instrumental variable strategy

In Table 15, we provide our estimates based on applying an instrumental variable estimation, for which we use variation in a product's exogenous property of network effects. For this, we use (1) the presence of users' interaction ($D_{UsersInteract}$), (2) the assessment, whether other users are necessary when using the app on a scale from 0 to 1 (Users' NE Assessment) and (3) the presence of keywords suggesting network effects ($D_{NEKeywords}$). These three measures shall instrument the HHI of the app's market and its market share. Table 15 shows the results of the IV in the cross section and accordingly the first stages for both endogenous regressors as the dependent variable and our primary demand measure based on the number of ratings (columns 1 and 2). The results are qualitatively similar for our second demand measure. The coefficients of the instruments show that the presence of network effects is positively and significantly correlated with the market concentration an app is in (for all the three measures), whereas for the market share this positive relationship applies to the presence of user interaction and the assessment that other users are necessary for the app's use. Values for the F-Test of well above 10 suggest the instruments to be relevant. Using the instrumented HHI and market share, we run the same specification as before with the number of privacy-sensitive permissions as the dependent variable (column 3). The results show the sign and significance for the coefficients of interest HHI and $Log.\ Market\ Share$ to remain the same. However, the coefficients are distinctively larger than in the baseline.

Table 15: Network effects as instrumental variables

	Log. Market Share	HHI	$\#_{DataCollection}$
	CS1	CS1	CS1
$D_{UsersInteract}$	0.297***	0.003***	
	(0.012)	(0.001)	
Users' NE Assessment	1.202***	0.080***	
	(0.021)	(0.001)	
$D_{NEKeywords}$	-0.133***	0.003***	
	(0.007)	(0.000)	
HHI			4.856***
			(0.307)
Log. Market Share			0.153***
			(0.019)
Log. Ratings	0.713***	-0.003***	-0.178***
	(0.002)	(0.000)	(0.014)
D_{Paid}	-0.798***	-0.026***	-0.188***
	(0.014)	(0.001)	(0.011)
$\#_{CleanPerms.}$	0.033***	0.001***	0.313***
	(0.001)	(0.000)	(0.002)
$D_{InAppProduct}$	-0.465***	0.003***	-0.132***
	(0.011)	(0.001)	(0.011)
Avg. Rating	0.108***	-0.002***	-0.015***
	(0.004)	(0.000)	(0.004)
Log. Length Description	-0.041***	-0.008***	0.021***
	(0.003)	(0.000)	(0.003)
$D_{IncludesAds}$	-0.692***	-0.015***	-0.241***
	(0.007)	(0.000)	(0.010)
Content Rating	0.185***	-0.000	-0.008**
	(0.004)	(0.000)	(0.004)
$D_{PrivacyProfile}$	-0.022***	-0.003***	0.399***
	(0.007)	(0.000)	(0.005)
$\#_{AppsbyDeveloper}$	-0.000***	-0.000***	0.000***
	(0.000)	(0.000)	(0.000)
App Age	0.000***	0.000***	0.000***
	(0.000)	(0.000)	(0.000)
Constant	-11.137***	0.215***	0.303
	(0.036)	(0.002)	(0.270)
Mean $\#_{DataCollection}$			1.324
Observations	783,807	783,807	783,807
F-Test 1st Stage			1,554.75
Centered R^2			0.20

Notes: In this table, we provide instrumental variable estimations. We use three instruments for the HHI and the market share: (1) the presence of users' interaction ($D_{UsersInteract}$), (2) the assessment, whether other users are necessary when using the app on a scale from 0 to 1 (Users' NE Assessment), and (3) the presence of keywords suggesting network effects ($D_{NEKeywords}$). In columns 1 and 2, we show the first stages for both endogenous regressors as the dependent variable and our first demand measure. Column 3 contains the second stage results again with the number of privacy-sensitive permissions as the dependent variable. Heteroscedasticity-robust standard errors in parentheses: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Download ZEW Discussion Papers from our ftp server:

http://ftp.zew.de/pub/zew-docs/dp/

or see:

https://www.ssrn.com/link/ZEW-Ctr-Euro-Econ-Research.html
https://ideas.repec.org/s/zbw/zewdip.html

//

IMPRINT

ZEW – Leibniz-Zentrum für Europäische
Wirtschaftsforschung GmbH Mannheim
ZEW – Leibniz Centre for European
Economic Research

L 7,1 · 68161 Mannheim · Germany
Phone +49 621 1235-01
info@zew.de · zew.de

Discussion Papers are intended to make results of ZEW research promptly available to other economists in order to encourage discussion and suggestions for revisions. The authors are solely responsible for the contents which do not necessarily represent the opinion of the ZEW.