Access to User Data, Market Power and Innovation in Online Markets: Evidence from the Mobile App Industry

Reinhold Kesler*
Centre for European Economic Research (ZEW)

Michael Kummer†
Georgia Institute of Technology & Centre for European Economic Research (ZEW)

Patrick Schulte‡
Centre for European Economic Research (ZEW)

January 17, 2017

PRELIMINARY - PLEASE DO NOT CIRCULATE

Abstract

We study how developers’ access to user data mediates the relationship between market power and product innovation in the mobile app industry. Developers which have high market power (1) might have more access to user data, both through active data collection, which might imply less user privacy, and through data provision by users. These data might then allow developers to generate more successful product innovations. We provide evidence on these questions using data on nearly 2 million apps from Google’s Play Store which we obtained on a quarterly basis in 2015 and 2016. We augmented these data with information on apps’ privacy-intrusiveness taken from PrivacyGrade.org as well as with information from AppBrain.com covering additional innovation measures and information on code libraries used by apps. First results suggest both a positive relationship between (1) market power and data access and (2) between data access and innovation.

JEL Classification: D12, D22, L15, L86

Keywords: User data; privacy; market power; product innovation; app markets.
1 Introduction

The shift of economic activity from offline to online markets has opened up many opportunities for large welfare gains. However, the availability of (big) user data and other typical features of such markets, go hand in hand with industry dynamics that raise important economic questions. Companies like Google, eBay, Amazon and other well-known firms active in online markets appear to have considerable market power. As a consequence, these firms’ conduct increasingly raises suspicion by competition authorities and consumer protection organizations alike. However, aside from anecdotal evidence there is little evidence with respect to the effects of market power on innovation and on privacy. This gap in the literature is surprising, because both innovation and a sufficient level of privacy are important long-run goals of economic policy. Innovation is a key determinant of long-run economic growth and the protection of privacy is a basic human right. Thus, understanding how market power affects innovation or privacy, and charting out the potential trade-off between these two outcomes is key for devising successful regulatory policies.

In this paper, we close this gap and provide theory-based empirical evidence on the following main research questions: (1) What is the relationship between firms’ market power and their innovation activity in online markets? (2) Is the availability of user data a relevant channel which helps explaining the competition-innovation relationship in online markets? With respect to the second question we ask: (i) Does more market power allow firms to collect and receive more user data, i.e. does more market power imply less privacy for users? (ii) Does access to more data allow firms then to develop and introduce more innovations and innovations which are better aligned with users’ preferences?

To provide such microeconometric evidence, we study data from the mobile app industry. This industry is highly innovative and economically relevant with a turnover of around US$ 41 billion in 2015 (AppAnnie, 2016), which can be considered a representative and increasingly important example of online markets. Our analysis is grounded on the methods of the long-standing theory-based empirical literature on competition and innovation (Aghion et al., 2005; Cohen, 2010). We base our research on theoretical models on innovation, such as the knowledge production function approach or the Crépon-Duguet-Mairesse (CDM) model (Crepon et al., 1998), where innovation is modeled as a (nonlinear) function of innovation inputs and market power.

For our empirical analysis, we exploit a unique and innovative data base, which contains
for 2015 and 2016 quarterly product-level information not only for a subset but for nearly the full set of apps available in Google’s Play Store (up to 2 million apps). The information covers detailed information about the apps, their developers and competitors. It allows constructing precise and innovative app-specific measures of innovation, market power and developers’ data access and thus lends itself perfectly to studying our research questions. We consider the rich information contained in our data set a drastic improvement over typical data limitations in the existing literature and thus see it as an enormous research opportunity. We apply various microeconometric estimation methods, such as simple OLS, Probit models and GMM-methods. To tackle the challenges in identifying causal effects we will use exogenous variation in market power. This variation stems from Google’s random promotional activities, potential policy shocks and unexpected market entries. We will apply both a difference-in-differences approach and IV-methods to exploit this information.

Combining our data with this empirical strategy, we show how market power is related to innovation in the online market for mobile applications. Moreover, we highlight how this relationship is shaped by the availability of user data. In addition, we will provide first empirical evidence on the relationship between firms’ market power and its consequences for users’ privacy.

These findings are valuable from a scientific, political but also managerial perspective. We provide new evidence on the key relationship between market power and innovation, which, so far, was mainly studied in offline markets. Moreover, we improve the understanding of the general mechanisms underlying the competition-innovation relationship by highlighting the role of user data. Finally, our research also contributes to the ongoing policy debate of about the consequences of market power (or competition) for innovation and privacy and the associated trade-offs. Thus our findings are relevant for policy makers, but also help platform owners to adjust platform policies to ensure optimal innovation and privacy levels in the long-run.

2 Literature

The effect of competition on innovation could be positive or negative. On the one hand, market power can increase firms’ ability (and incentives) to extract post-innovation rents (Schumpetarian effect) by preempting competition. On the other hand, innovation can replace monopoly profits (replacement effect) (Dasgupta and Stiglitz, 1980; Arrow, 1962). In the presence of less market power, profits are lower and thus the incentives to innovate are reduced as well. At the
same time firms might want to escape competition (escape-competition effect) through innovation (Aghion et al., 2005). As a result, several established theories on competition and innovation generate conflicting predictions. These predictions range from a monotonic relationship (both positive and negative) to an inverted u-shaped relationship (i.e. neither too little nor too much competition is good for innovation) by Aghion et al. (2005). Yet, empirical evidence remains ambiguous and is especially scarce with respect to digital markets.

A recent study focused on apps for personal digital assistants (PDAs) reveals that additional app developers of a similar type decrease innovation incentives, whereas for more developers of a different type they increase, with the former being the dominating factor (Boudreau, 2012). Boudreau and Jeppesen (2015) consider commercial game engine platforms and find that development rates of producers of complementary goods (complementors) decrease with more competitors, while they increase with growing platform usage. However, more complementors also increase the platform usage, which results overall in a zero effect. Using data from a platform of unofficial apps, Miric (2015) finds a negative impact of the number of competing products on the innovation rate, which is more pronounced for paid producers. In the context of smartphone applications, Liu et al. (2014) study how the threat of a competitor’s entry influences the timing and quality of cross-platform entry. Their results suggest a negative impact on the app quality due to a premature release. Furthermore, a recent study considers the effect of a platform owner entry on the complementors’ innovation behaviour and finds a positive impact, which is attributed to an increased attention (Foerderer et al., 2016). Besides new product releases and major updates, minor updates are also considered as measures of innovation. Exemplary studies analyse the effect of competition on the time to patch software vulnerabilities with conflicting evidence of a reduction in time due to more competitors (Arora et al., 2010) in contrast to a time increase (Jo, 2016).

The heterogeneous evidence concerning the competition-innovation relationship necessitates accounting for the characteristics of the respective market. For online markets, one distinctive feature is data access, which, at least to our knowledge, has not been considered so far by studies that look at the competition-innovation relationship in the digital markets context.

Two strands of literature study the relationship between competition and firms’ access to user data. One strand studies whether the amount of data firms actively collect, so called data collection, increases or decreases with their market power. In contrast, a second strand is concerned
with the data users provide, so called data provision, in dependence of firms’ market power.

Within the first strand, studying the data collection by firms, theoretical work presumes that market power comes on average with more data collection by firms (Casadesus-Masanell and Hervas-Drane, 2015), a result which is supported by Brown (2016), who reviews the literature on the economics of privacy. Data are assumed to be valuable for firms, be it as a direct means for revenue generation (selling the data to others), be it for targeted advertisement or be it for the implementation of user-specific pricing. At the same time, if firms collect data about users and users realize this, it typically reduces the product’s quality from the perspective of the user by being detrimental to its privacy (see e.g. Kummer and Schulte (2016)). Data can thus be considered a second currency firms will be able to collect to a higher extent the higher their market power. Empirical evidence, which only scarcely exists, supports this hypothesis. Preibusch and Bonneau (2013), using descriptive evidence for 140 web sites from five internet industries, find that websites having no major competitor collect significantly more data than those having competitors. Besides, there is, to our knowledge, no other empirical study providing evidence on the relationship between firms’ market power and their active data collection behaviour.

Whereas this first form of data is collected actively by firms, user generated data, e.g. product reviews, have to be provided by users. A second strand of the literature studies the relationship between market power and the availability of this form of information for firms. Gans et al. (2017) analyse the consequences of market power for users’ likelihood to ‘raise their voice’ in reaction to quality shocks in products. According to their model and their empirical evidence (studying tweets and the US airline industry), users complain relatively more often about quality shocks if the firm providing the respective service has relatively high market power. The reason is that such customers have less relevant alternatives to choose from and more to lose from leaving the firm’s service such that they choose to complain instead to leave. This argument might also hold not only for cases of negative quality shocks but might also hold for cases where customers have ideas for product improvements (innovations) in general. Thus, for firms such information might be a valuable asset, i.e. an additional source of information which allows them to react faster and to be more targeted to user expectations. In the case of apps, such information is contained in users’ explicit textual app reviews, which consist of shorter or longer reviews of a given product.

Taken together, there is reason to believe that more market power comes with more data for
firms, both in the form of collected and provided data, which might be a valuable asset that in a second step can be exploited by firms for improving their innovation performance. So far, very little empirical evidence exists with respect to this question. Especially, no evidence exists for app markets, which is one of the most important markets with respect to data collection and user privacy (Kummer and Schulte, 2016). We address this gap by studying this relationship more rigorously, analysing both forms of information, and comparing for both the relevance of the outlined theoretical results.

As described, two possible data sources are potentially available for app developers to enable data-driven innovation: data collection and data provision. Hence relevant literature strands analysing data-driven innovation comprise both studies arguing that the data access of firms is related to innovation as well as research on open innovation, which relates external knowledge to firms’ innovation.

Given its novelty, there are only a few empirical papers studying the hypothesis of data-driven firm success. The research questions range from studying data-driven decision making to investments in big data technologies (such as Hadoop). Their findings suggest a positive impact on firms’ productivity (Brynjolfsson et al., 2011; Brynjolfsson and McElheran, 2016; Tambe, 2014). Accordingly, these data-driven returns are augmented by complementary skills (Brynjolfsson and McElheran, 2016) or access thereof (Tambe, 2014) along with necessary IT and data assets (Brynjolfsson and McElheran, 2016; Tambe, 2014). Saunders and Tambe (2015) also find a positive impact of data-driven practices on market value and firm profitability by studying the extent to which data-related activities are mentioned in the company’s annual report from 1996 to 2012. Besides an overall increase in data activities and a dispersion from IT-producing to IT-using firms in this period, their results suggest that data-intensive firms outperform competitors due to their data assets. Despite first evidence of data-driven firm success, these studies primarily focus on specific industries or publicly traded companies, while additionally, at least to our knowledge, none of these consider the impact on innovation.

Following the open innovation paradigm by Chesbrough (2003), firms can realize innovation by sourcing ideas from outside the firm. The evidence from empirical studies suggests in general a positive impact on firms’ innovation outcomes from external knowledge sourcing (Greco et al., 2015). However, as West and Bogers (2014) argue, individuals as a source of innovation are rarely the subject of open innovation research. Most studies on external individuals’ involvement for
firms’ innovation look at the ideator’s motivation, the ideation process and the ideator’s outcome as well as its determinants (Schemmann et al., 2016). Bertschek and Kesler (2016) find a positive impact from the intensity of user feedback on companies’ Facebook Pages on the firm’s probability to realize a product innovation for German firm-level data. Hence, the literature on open innovation has hardly studied how customers contribute to the innovation process, specifically, to our knowledge, no evidence exists on whether product reviews by customers are a source of external knowledge helping firms to innovate.

Summarizing, even though there is a long-standing theoretical and empirical literature studying the relationship between competition and innovation in offline markets, there is only little evidence for online markets. A few very recent studies analyse the relationship between competition and innovation in online markets. However, they do not explicitly consider the important specific characteristic of online markets, i.e. the availability of user data as a source for data-driven innovation did not deserve sufficient attention. Also, convincing empirical evidence on the relationship between competition and user privacy is lacking. Evidence on these questions, however, would be of high value for competition policy, in order to understand the consequences of market power and the trade-offs which come along with it. We therefore aim to close these gaps by analyzing the role of data access and data-driven innovation as an important channel for this relationship. This analysis will also provide valuable insights on the consequences of firms’ market power for users’ privacy.

3 Empirical Framework

In this paper we want to shed light on one potential channel of the competition-innovation relationship in online markets, namely the availability of user data and the resulting data-driven innovation. The aim is first to understand whether more market power allows firms to access more data about users’ needs. This enhanced access to data reduces users’ privacy and aggravates the asymmetry in the markets. Secondly, we analyse whether these additional data, both collected and provided, allow firms then to develop and introduce more and higher-valued innovations.

To study the two-step relationship between (1) competition and data access as well as (2) data access and firms’ innovation output, we will build on the rich theoretical and microeconometric innovation literature and the methods applied therein. Mohnen et al. (2007) compare and discuss the approaches applied to study innovation. A model very often applied to study firms’ innovation
is the Crépon-Duguet-Mairesse (CDM) model developed by Crepon et al. (1998), which models the determinants of innovation inputs and outputs in several steps. They explain innovation success by the use of innovation inputs such as R&D investments. We extend the notion of innovation inputs by studying the role of user data. Specifically, rather than studying the role of market power for firms’ R&D investments and subsequent innovation, we study the role of market power for data access and the subsequent innovation success.

For the first step, the theoretical literature predicts a positive relationship between market power and data collection by firms with data serving as a second currency (Casadesus-Masanell and Hervas-Drane, 2015; Brown, 2016). Empirical studies in this field are scarce and are urgently needed. Additionally, Gans et al. (2017) show that customers of firms with greater market power ‘raise their voice’ in the presence of quality shocks more often, and thus provide additional data.

For the second step, empirical studies show evidence of data-driven firm success, especially with respect to firms’ productivity, but they often have selective samples and do not consider innovation as an outcome variable (Brynjolfsson et al., 2011; Brynjolfsson and McElheran, 2016; Tambe, 2014; Saunders and Tambe, 2015). In the context of open innovation, there is evidence of a positive impact from sourcing external knowledge on firms’ innovation, while typically consumers (and especially user reviews) are not considered as a source (Greco et al., 2015; West and Bogers, 2014).

Our baseline empirical strategy to test the predicted two-step relationship can be illustrated by the following two simple equations:

\[
\text{Data}_{it} = \alpha + \beta_1 \text{MP}_{it} + \text{CONTROLS} + \epsilon_{it} \\
\text{Inno}_{it} = \alpha + \beta_2 \text{Data}_{it} + \text{CONTROLS} + \epsilon_{it}
\]

where \(\text{Inno}_{it}\) and \(\text{MP}_{it}\) are measures of innovation and competition of a given product \(i\) at time \(t\) and \(\epsilon_{it}\) is an stochastic error term. We can use two measures to quantify \(\text{Data}_{it}\): 1) data collection practices of a given firm and 2) data from users’ product reviews.

To identify the causal effect of competition on innovation and to analyse the role of data access as a possible driver of the competition-innovation relationship, obviously, we have to bear in mind that there are identification challenges. E.g. innovation can increase a firm’s post-innovation market power which implies the existence of reverse causality. Thus, besides
employing our baseline microeconometric (panel) approaches such as OLS, Fixed-Effects or Probit estimation, we plan to use sources of exogenous variation in competition intensity and to exploit this information by applying a difference-in-differences approach and IV-methods.

Our main source of exogenous variation stems from Google’s promotional activity named “Deal of the Week”, which promotes two apps for one week in several countries. These apps are promoted during this week prominently throughout the Google Play Store, are offered for a significantly lower price (equal to 10 Cents) and, are chosen by Google, such that developers have little influence on its choice. The promotion can be expected to lead to an increased demand for the promoted app and to a relative decrease in the competitors’ demand. Indeed, studies concerning such large-scale promotions in app markets, which looked at the effect on sales and ratings (Askalidis, 2015; Chaudhari, 2015), suggest that such promotions can affect competitors’ sales and that this effect can last beyond the promotional period. This implies that there is a long-run effect on the relative demand for such affected apps, which represents an exogenous shock to their market power. Alternatively, we plan to exploit app-specific variation due to unexpected market entries of new apps. Often the popularity of new apps is hard to predict for incumbents and for the developers of the new apps, such that their effect on the market power of incumbents can be considered exogenous. An example could here be the entry and success of the Pokémon GO app, whose success came unexpected and which affected the demand of similar games strongly. A second alternative we have in mind is an exogenous shock to the competition intensity of certain apps which soon could arise from the ruling of the European Commission (EC) concerning the pre-installed apps that are required by Google to be on mobile devices with the Android operating system.¹ If the decision by the EC is to weaken these requirements of pre-installation, it would affect corresponding app categories, in which Google apps have been pre-installed previously (e.g. search or browsing applications). As the respective Google app loses its market barrier, competitors have a greater exposure to consumers and thus can potentially gain a higher market power.

These exogenous shocks enable us to identify both equations of interest. On the one hand, the relationship between market power and data access can be inferred from this variation. On the other hand, once the relationship between data access and market power is established, the exogenous variation in the competition intensity may also serve as an instrument when studying

the determinants of innovation.

Furthermore, by considering the chronological sequence of textual links between user reviews and the description of updates, we can mitigate reverse causality and effectively identify the causal direction of the effect.

Finally, the robustness of the results can be tested across varying measures of competition, innovation data collection practices of firms and user reviews, at other levels of aggregation such as categories (similar to an industry-level approach), as well as distinguished by the market position of the product/developer (leader/laggard) and by the developer characteristics (e.g. size/experience).

4 Data

To implement the outlined empirical model, we have compiled a unique and innovative data base about the app market at the product-level, which will be exploited for this paper. The data set covers nearly all apps available in Google’s Play Store (up to 2 million apps) and contains quarterly product-level information for 2015 and 2016 allowing analyses in panel format. The data contain detailed information about the apps, their developers and their competitors. In the following, we describe the data base and how we can exploit it for our research project.

4.1 Data Description

In a first step we describe our raw data taken from Google’s Play Store and following that outline the variables of interest we derive from the raw data.

Raw Data

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

- the total number of installations of an app,
- (monthly) downloads of an app,
- information on updates (date, textual information on what is new, version number)
- the names and IDs of similar apps
the permissions which an app requests upon installation and which apps require to perform certain functions (around 140 such permissions exist, which include e.g. ‘network access’, ‘read contents of USB’, ‘read contact data’, ‘read browser data’, ‘read sensitive log data’),

number and values of quantitative ratings (from 1 to 5 stars),

is the app an editor’s choice (yes/no)

textual reviews (date, rating from 1 to 5 stars, content, availability of a developer-response)

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.),

code size (in KB),

apps’ description (length, content) and its illustration in the Play Store (video and screenshot availability),

content rating (USKs),

availability of interactive elements (e.g. ‘users interact’, ‘digital purchases’ etc.),

Android version required for installation,

In addition, the data also contain developer-specific information on:

the name of the developer,

top developer status (yes/no),

number of its apps,

the set of its available apps.
Innovation

In empirical studies analysing product innovation, the innovation success is typically measured either by patents or by survey responses indicating whether firms have introduced a new product innovation during the survey period (Cohen and Levin, 1989; OECD and Eurostat, 2005; Cohen, 2010). Studies applying web-scraped data sets, in contrast, offer additional scope to measure innovation. In our case, we have detailed information on product updates and new product releases. The information on such events allows us not only to determine whether there has been an innovation or not but provides us with much deeper insights about the properties of these innovations. New app releases, for example, can be considered as a form of radical innovation. For such releases we know the release date and have detailed information about the apps’ properties, such as the apps’ contents, its functions, its reviews etc. and know the market it is released into. For updates, which can be considered incremental innovations, we get a textual description of ‘what is new’ (from a section containing the verbal listing of all changes coming with the most recent update), can retrieve the difference in the apps’ description (content and length), observe the change in the version number, see how the app’s code size has changed and can even identify changes in apps’ permissions (which apps’ have to ask for in order to perform certain functions). It is obvious that by exploiting this information using, among others, linguistic methods, we can measure innovation in an unprecedented way. We are able to identify in a very detailed way what has changed, i.e. the content of an innovation. It also allows us to distinguish between minor and major updates, i.e. between updates providing only smaller bug fixes and updates which introduce new functionality. Finally, since we observe product reviews (quantitative ratings but also qualitative, verbal reviews) before and after the innovation, we can study how users assess the respective innovation, which can be used to measure innovation quality.

The information on the exact release date of an app is currently retrieved from the platform AppBrain.com. AppBrain.com also provides a more detailed changelog with respect to updates, which will serve as the basis for a measure of innovation frequency. Moreover, together with a one-year history of the average rating collected by AppBrain.com this changelog enables a more precise before and after quality measure of updates.

2Web-scraped data not only offers alternative measures of innovation, applying standard measures such as patent data here would also not be very reasonable, since, as e.g. Boudreau et al. (2015) note, only a very small fraction of app developers actually file patents, such that patents would not be a representative measure of innovation in such a market.
Taken together, in contrast to standard data sets, our data allow us to study not only the binary existence of an innovation but allow us to study for both more radical and more incremental innovations the innovation frequency, the innovation content and the innovation quality. We want to stress that we consider this a drastic improvement over typical data limitations in the existing literature.

**Market Power**

In online markets, one major difficulty is to define a market and to measure market power, because traditional approaches are typically hard to apply. A distinctive feature and big advantage of our data set, however, is the availability of information on app-specific competitor apps. Google’s Play Store nowadays provides for each app information on a set of “similar apps”. These apps are selected according to their similarity in functionality. For each app, between 0 and 24 of such competitors are presented in the Play Store (the actual set of competitor apps, however, is in several cases even bigger). We can use the IDs and names of the apps from this set to identify those competitors in our data set (which covers the full population of available apps). Using the information we possess about the competitors, we can construct various detailed, app-specific measures of competition and market power. E.g., we can compute for each app a market share or a market concentration measure by exploiting the information about the relevant competitors and the information on their individual download numbers and ratings. Being able to construct such detailed, app-specific measures is a big advantage over alternative data sets, which typically only have information on the number of competitors or even worse only have industry-specific competition measures available. Of course, as a robustness check and for comparability, we can construct such basic measures, too. E.g., we can use the app-specific number of competitors or we can construct category- or subcategory-specific competition measures, similar like those in studies only having industry-level competition measures at hand. The precise information on the entry and exit of apps is currently retrieved from the platform AppBrain.com enabling to study unexpected market dynamics.

Thus, our data set provides an exceptionally rich source for studying empirically the consequences of competition and market power. It will not only allow for deep insights into the competition-innovation relationship but will also allow for methodological insights by allowing to compare several novel and more traditional competition measures and that way providing evidence on their relevance for competition economics.
Data Access

In a next step we construct measures of firms’ data access to allow analysing data-driven innovation as a mediator of the competition-innovation relationship. As sketched in section 3, we will compare two sources of information which are potentially available to firms: 1) ‘actively’ collected data (using the permissions an app developer can request upon installation) and 2) ‘passively’ received user comments (short or longer textual publicly available reviews).

Data Collection

The amount and variety of information app developers can collect about their customers through their app strongly depends on the permissions the app requests upon installation from the user. Such permissions can e.g. allow an app to ‘read contact data’, to ‘read browser data’, to ‘read sensitive log data’ etc. Overall, there are around 140 different permissions available from which a developer can choose and which he can ask for. Some of these permissions allow the developers to collect information about users’ behaviour and preferences and can therefore be considered as privacy-intrusive. Kummer and Schulte (2016) based on various classifications (e.g. Sarma et al., 2012) identify 12 permissions as privacy-sensitive and at the same time as informative for the app developer. These 12 permissions include among others the following permissions: ‘read phone state and ID’, ‘fine gps location’, ‘read sms or mms’, ‘read contact data’, ‘read browser data’ and ‘read sensitive log data’. This selection of permissions can be a starting point for identifying all permissions allowing developers to collect information about users’ app usage behaviour, related problems and wishes to improve the apps. However, at the same time, it will be an important task to identify all the relevant permissions for which we will e.g. referencing computer scientists’ work regarding such issues.

An additional source of information on apps’ ability to collect information about users we exploit comes from PrivacyGrade.org, which is the so far most comprehensive effort of computer scientists to evaluate the privacy-intrusiveness of apps. Lin et al. (2012, 2014) analyzed in 2014 and 2016 detailed information about more than one million apps’ privacy-related behavior. They summarize their findings in the form of a grade, ranging from A+ (most privacy sensitive) to D (least privacy sensitive). Grades were assigned using a privacy model that measures the gap between people’s expectations of an app’s behavior and the app’s actual behavior.

3Their results are provided publicly: see: http://privacygrade.org/
The information on third-party libraries\textsuperscript{4} used by the app is retrieved from the platform AppBrain.com and allows further analysis of data collection practices in connection with used third-party libraries. For example, the information provided by AppBrain.com allows distinguishing between ad-related and social libraries as well as developer tools but also allows for own classifications of code libraries.

**Data Provision**

A second source of information app developers can use are users’ textual reviews. Users have the option to provide a simple quantitative review consisting only of a rating (1 to 5 stars can be given for an app) or to write a smaller or bigger review of the app. In such a case the app developer also has the option to reply publicly to a published review. The reviews can be used to criticize certain aspects of an app or to express suggestions for improvement. We can use these reviews (their quantity but also their content) directly as a measure of user-provided information available to developers. In a further step, we will apply linguistic methods to analyse the content of those reviews (using so-called topic-specific dictionaries), which allows measuring the reviews’ sentiment and whether they contain requests or feedback.\textsuperscript{5} Having identified the content of these reviews also allows drawing a direct link between user reviews and subsequent innovation. For doing so, we will study whether specific contents of users’ textual reviews can be found subsequently in the descriptions of the app updates. For this, the appearance of the same keywords or combinations thereof is analysed (e.g. whether specific errors or permissions are mentioned). Both sources of information, the collected information through permissions and the information provided by reviews, can be studied either separately or in a combined index.

**4.2 Descriptive Evidence**

In the Appendix we provide first descriptive evidence. Table 1 contains both for the cross-section from September 2016 and for the balanced panel summary statistics of the main variables of interest as well as of some control variables. Figures 1, 2 and 3 provide histograms illustrating the distribution of our main innovation, data access and competition variables.

\textsuperscript{4}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 \url{http://privacygrade.org/third_party_libraries} for an exemplary overview.)

\textsuperscript{5}See McIlroy et al. (2016) for studies on how to extract information from user reviews.
5 Estimation Results

Table 2 contains simple correlations regarding the relationship between market power and data access. Columns 1 to 3 provide results regarding apps’ data collection behavior, whereas columns 4 and 5 contain evidence regarding the data provision by users. Specifications 2 and 5 provide fixed-effect panel-based evidence, whereas the remaining 3 specifications contain cross-section based evidence. In all specifications the variable of interest is the apps’ log. market share. In specification 1 and 2 the dependent variable is a dummy which is equal to one if the app has at least one privacy-sensitive permission ($D_{PrivPerm}$). In specification 3 the dependent variable is the log number of app libraries an app uses (information stems from AppBrain.com). In columns 4 and 5 the dependent variable is our measure of data provision by users, i.e. the log number of explicit reviews in the last 3 months. These first results are therefore in line with the idea of market power resulting in higher data access for app developers.

To provide first evidence whether these additional information allow app developers to improve their apps through more and larger updates, we analyze in table 3 the relationship between data access and app innovation. The table contains simple correlations regarding the relationship between data access of apps and apps’ update behavior. Columns 1 to 3 provide evidence regarding apps’ update frequency whereas columns 4 to 6 contain evidence regarding the size of updates. Specifications 2, 3, 5 and 6 contain fixed-effect panel based evidence whereas specifications 1 and 3 provide cross-sectional evidence. In specification 1, 2 and 3 the dependent variable is an dummy which is equal to one if the app has been updated within the last 3 months ($D_{Update}$). In the remaining specifications the dependent variable is the size of the ‘what is new’-section. In specification 1 and 4 the independent variable is the log number of app libraries, in specifications 2 and 5 it is a dummy which is equal to one if the app has at least one privacy-sensitive permission. In specification 3 and 6 the dependent variable is the log number of explicit user reviews within the last 3 months. In all specifications the we control for the log. market share of an app. The results again indicate to a positive relationship between data access and innovation. Both for innovation size and for innovation frequency these results as well as for both data provision and data collection there seems to be a positive relationship.
Limitations and Further Research

The findings we presented above are only first results. However, they showcase the potential of our data and suggest that access to user data might indeed explain an important part of the relationship between market power and innovation in online markets.

Further research should attempt making use of exogenous variation from Google’s promotional activity (named “Deal of the Week”), or from unpredictable market entry of new apps, the popularity of which is hard to predict for incumbents (and for the developers of the new apps alike). A second alternative we have in mind is an exogenous shock which soon could arise from the the European Commission’s ruling about Google’s pre-installed apps. Incorporating such exogenous shocks would enable us to better identify both the relationship between market power and data access the determinants of innovation.

Furthermore, by considering the chronological sequence of textual links between user reviews and the description of updates, we can mitigate reverse causality and effectively identify the causal direction of the effect. Finally, the robustness of the results can be tested across varying measures of competition, innovation and data collection practices, and at other levels of aggregation.

6 Conclusion

In this paper we argue that developers’ access to user data affects the relationship between market power and product innovation in the mobile app industry. Developers which have high market power (1) might have more access to user data, both through active data collection, which might imply less user privacy, and through data provision by users, which together (2) might allow developers to generate more and more successful product innovations. To provide evidence on this relationship we use data from nearly 2 million apps from Google’s Play Store which we obtained on a quarterly basis in 2015 and 2016. We augmented these data with information on apps’ privacy-intrusiveness taken from PrivacyGrade.org as well as with information from AppBrain.com covering additional innovation measures and information on code libraries used by apps.

Our first results are in line with the idea of a positive relationship between market power and data access. Apps having a higher market share seem to have both more access to data via data collection and data provision. At the same time, apps which have more access to data also
show a higher innovation activity, i.e. they are more likely to be updated and their updates are of a larger size. These results are based on cross-sectional and panel evidence, and provide insight on the conditional correlations in the data. In future steps we plan to improve our identification strategy and exploit exogenous variation that allows for a causal interpretation of our findings. To do so we will implement a more rigorous design which exploits information on unexpected market entries of rivals, using information on promotional activities by Google as well as by exploiting policy changes implemented by Google.
References


## Appendix

### Table 1: Summary Statistics of Cross-Section and Panel

<table>
<thead>
<tr>
<th></th>
<th>Cross Section</th>
<th></th>
<th>Balanced Panel</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mean</td>
<td>p50</td>
<td>min</td>
<td>max</td>
</tr>
<tr>
<td>update_age</td>
<td>524.24</td>
<td>420.00</td>
<td>0</td>
<td>2904</td>
</tr>
<tr>
<td>$D_{Update}$</td>
<td>0.15</td>
<td>0.00</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>maj_upd_dum</td>
<td>0.07</td>
<td>0.00</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Size of What is New</td>
<td>1.62</td>
<td>1.00</td>
<td>0</td>
<td>58</td>
</tr>
<tr>
<td>num_pp2</td>
<td>1.25</td>
<td>1.00</td>
<td>0</td>
<td>12</td>
</tr>
<tr>
<td>$D_{PrivPerm.}$</td>
<td>0.53</td>
<td>1.00</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>dnr_exp</td>
<td>10.62</td>
<td>0.00</td>
<td>0</td>
<td>4044</td>
</tr>
<tr>
<td>num_comp_dnr_nod</td>
<td>19.98</td>
<td>16.00</td>
<td>4</td>
<td>120</td>
</tr>
<tr>
<td>msh_dnr_nod</td>
<td>0.02</td>
<td>0.00</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>hhi_dnr_nod</td>
<td>0.33</td>
<td>0.26</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>price</td>
<td>0.23</td>
<td>0.00</td>
<td>0</td>
<td>350</td>
</tr>
<tr>
<td>num_installs</td>
<td>113.92</td>
<td>3.00</td>
<td>0</td>
<td>75000</td>
</tr>
<tr>
<td>inapp</td>
<td>0.11</td>
<td>0.00</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>avgrat</td>
<td>3.99</td>
<td>4.10</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>num_screenshots</td>
<td>7.52</td>
<td>6.00</td>
<td>0</td>
<td>32</td>
</tr>
<tr>
<td>video</td>
<td>0.12</td>
<td>0.00</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>desc_length_en</td>
<td>1126.67</td>
<td>808.00</td>
<td>1</td>
<td>11922</td>
</tr>
<tr>
<td>size_KB</td>
<td>12.30</td>
<td>5.22</td>
<td>0</td>
<td>3670</td>
</tr>
<tr>
<td>num_by_dev</td>
<td>53.83</td>
<td>9.00</td>
<td>1</td>
<td>2490</td>
</tr>
<tr>
<td>N</td>
<td>563815</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure 1: Histograms of Innovation Variables

(a) Innovation Frequency (Update Age)
(b) Innovation Size (Size of What is New-Section)

Notes: The figures show, based on the cross-section from September 2016, histograms for two innovation measures. The left graph shows the distribution of apps’ update age, i.e. the days since the last update. The right graph shows the distribution of the sizes of the ‘what is new’-section of each app.
Figure 2: Histograms of Data Collection and Data Provision

(a) Data Collection (Number of Permissions)

(b) Data Collection (Privacy Grade)

(c) Data Collection (Number of App Libraries)

(d) Data Provision (Number of Verbal Reviews)

Notes: The figures show, based on the cross-section from September 2016, histograms for four measures of data access. The first graph shows the distribution of the number of permissions apps request. The second graph shows the distribution of the privacy grades (obtained from PrivacyGrade.org). The third graph shows the distribution of the number of app libraries an app uses (taken from AppBrain.com). The fourth graph shows the distribution of the number of new explicit textual reviews an app has received within the last three months.
Notes: The figures show, based on the cross-section from September 2016, histograms for three measures of apps’ market power and competitive situation. The first graph shows the distribution of the number of competitors of an app. The second graph shows the distribution of the log market share (computed using the number of additional ratings as a demand measure). The third graph shows the distribution of the app-specific Herfindahl-Hirschman-Index.
Table 2: Market Power and Data Access

<table>
<thead>
<tr>
<th></th>
<th>Data Collection</th>
<th>Data Provision</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
</tr>
<tr>
<td>Log. Market Share</td>
<td>0.010***</td>
<td>0.001***</td>
</tr>
<tr>
<td></td>
<td>(0.000)</td>
<td>(0.000)</td>
</tr>
<tr>
<td>Constant</td>
<td>0.529***</td>
<td>0.529***</td>
</tr>
<tr>
<td></td>
<td>(0.007)</td>
<td>(0.001)</td>
</tr>
<tr>
<td>Category</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Controls</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Observations</td>
<td>454633</td>
<td>447221</td>
</tr>
<tr>
<td>Num. of Groups</td>
<td>242207</td>
<td>242207</td>
</tr>
<tr>
<td>Adjusted $R^2$</td>
<td>0.06</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Notes: The table contains simple correlations regarding the relationship between market power and data access. Columns 1 to 3 provide results regarding apps’ data collection behavior, whereas columns 4 and 5 contain evidence regarding the data provision by users. In all specifications the variable of interest is the apps’ log. market share. In specification 1 and 2 the dependent variable is a dummy which is equal to one if the app has at least one privacy-sensitive permission ($D_{PrivPerm}$). In specification 3 the dependent variable is the log number of app libraries an app uses (information stems from AppBrain.com). In columns 4 and 5 the dependent variable is our measure of data provision by users, i.e. the log number of explicit reviews in the last 3 months. Specifications 2 and 5 provide fixed-effect panel-based evidence, whereas the remaining 3 specifications contain cross-section based evidence. Standard errors in parentheses: *** p<0.01, ** p<0.05, * p<0.1.
### Table 3: Data Access and Innovation

<table>
<thead>
<tr>
<th></th>
<th>Update Frequency ((D_{\text{Update}}))</th>
<th>Update Size (Size of What is New)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
</tr>
<tr>
<td>Log. App Libraries</td>
<td>0.086***</td>
<td>0.155***</td>
</tr>
<tr>
<td></td>
<td>(0.001)</td>
<td>(0.008)</td>
</tr>
<tr>
<td>Log. Market Share</td>
<td>0.033***</td>
<td>0.010***</td>
</tr>
<tr>
<td></td>
<td>(0.000)</td>
<td>(0.001)</td>
</tr>
<tr>
<td>(D_{\text{PrivPerm.}})</td>
<td></td>
<td>-0.016</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.010)</td>
</tr>
<tr>
<td>Log. Verbal Reviews</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>0.176***</td>
<td>0.256***</td>
</tr>
<tr>
<td></td>
<td>(0.008)</td>
<td>(0.006)</td>
</tr>
<tr>
<td>Category</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Controls</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Observations</td>
<td>278704</td>
<td>447221</td>
</tr>
<tr>
<td>Num. of Groups</td>
<td>242207</td>
<td>155011</td>
</tr>
<tr>
<td>Adjusted R²</td>
<td>0.09</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Notes: The table contains simple correlations regarding the relationship between data access of apps and apps’ update behavior. Columns 1 to 3 provide evidence regarding apps’ update frequency whereas columns 4 to 6 contain evidence regarding the size of updates. In specification 1, 2 and 3 the dependent variable is a dummy which is equal to one if the app has been updated within the last 3 months \((D_{\text{Update}})\). In the remaining specifications the dependent variable is the size of the ‘what is new’-section. In specification 1 and 4 the independent variable is the log number of app libraries, in specifications 2 and 5 it is a dummy which is equal to one if the app has at least one privacy-sensitive permission. In specification 3 and 6 the dependent variable is the log number of explicit user reviews within the last 3 months. Specifications 2, 3, 5 and 6 contain fixed-effect panel based evidence whereas specifications 1 and 3 provide cross-sectional evidence. In all specifications the we control for the log. market share of an app. Standard errors in parentheses: *** p<0.01, ** p<0.05, * p<0.1.