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)

June 8, 2017

PRELIMINARY - DO NOT QUOTE WITHOUT PERMISSION FROM THE AUTHORS

Abstract

We study how developers' access to user data mediates the relationship between market power and product innovation in the mobile app industry. Developers with market power might more easily access user data through active data collection, which might imply less user privacy. Better access to user data might then facilitate more successful product innovations. Our empirical evidence is based on data on nearly 2 million apps from Google's Play Store which we obtained quarterly in 2015 and 2016. We augment these data with information on apps' privacy-intrusiveness (from PrivacyGrade.org), and on app-specific innovation activity and usage of code libraries (from AppBrain.com). 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.

---

*We are grateful to Sinan Aral, Steve Berry, Irene Bertschek, Avi Goldfarb, Shane Greenstein, Maude Hasbi, Martin Peitz, Philipp Schmidt-Dengler, Christiaan Hogendorn, Yossi Spiegel, Frank Verboven, Mike Ward, Christine Zulehner as well as participants of the seminars at ZEW and GeorgiaTech, the MACCI Annual Conference 2017, as well as the IIOC 2017 for their helpful comments.

†P.O. Box 103443, D–68034 Mannheim, Germany. Email: `kesler@zew.de`.

‡221, Bobby Dodd Way, Atlanta, GA, 30308, USA. Email: `michael.kummer@econ.gatech.edu`.

§P.O. Box 103443, D–68034 Mannheim, Germany. Email: `schulte@zew.de`.

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 try to close this gap and provide theory-based empirical evidence on the following questions: (1) How is a firm's market power related to their innovative activity in online markets? (2) Is user data a relevant channel for explaining the relationship between competition and innovation in online markets? With respect to the second question we ask: (i) Does more market power allow firms to collect 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 answer these questions, 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 quarterly product-level information for 2015 and 2016. Our analysis exploits not only for a subset but almost the entire set of apps that were 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 richness of 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 estimation methods, and will use exogenous variation in market power to tackle the challenges in identifying causal effects. The source of exogenous variation is Google's random promotional activities, potential policy shocks and unexpected market entries. An independent source of identification will exploit a market's 'natural' intensity of network effects and the resulting tendency for higher or lower market concentration. 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 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 (Schumpeterian 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 innova-

tion (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.
One particular strand of literature studies the relationship between competition and firms' access to user data, i.e. whether the amount of data firms actively collect, so-called data collection, increases or decreases with their market power.

Within this strand, 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.

Therefore, there is reason to believe that more market power comes with more collected data by firms, 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 and comparing the relevance of the outlined theoretical results.

As described, collected data may potentially serve app developers to enable data-driven innovation. Hence, relevant literature strands analysing data-driven innovation comprise studies arguing that the data access of firms is related to 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.

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 explore the role of user data as a mediating channel in the competition-innovation relationship in online markets. Specifically, we are interested in how the availability of user data interacts with data-driven innovation. To analyze this question we proceed in two steps: First, we aim to understand whether more market power allows firms to access more data about users' needs. Such enhanced access to data might reduce users' privacy and would aggravate the asymmetry in the markets. In the second step, we analyze whether their advantage in accessing user data allows firms to subsequently 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, theoretical studies predict a positive relationship between market power and firms' data collection with data serving as a second currency (Casadesus-Masanell and Hervas-Drane, 2015; Brown, 2016). Empirical studies in this field are scarce and urgently needed.

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).

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

$$Data_{it} = \alpha + \beta_1 MP_{it} + CONTROLS + \epsilon_{it} \quad (1)$$

$$Inno_{it} = \alpha + \beta_2 Data_{it} + CONTROLS + \epsilon_{it} \quad (2)$$

where $Inno_{it}$ and MP_{it} are measures of innovation and competition of a given product i at time t and ϵ_{it} is a stochastic error term. We can use several measures to quantify $Data_{it}$ as data collection practices of a given firm.

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 and Fixed Effects 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.

Exogeneous Variation in Market Structure: 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. The popularity of new apps is often hard to predict for both incumbents and the developers of the new apps themselves. Thus, their effect on the market power of incumbents can be considered exogenous. An example was the entry and success of Pokémon GO, whose unexpected success affected the demand of similar games strongly. As a second alternative exogenous shock we consider an expected ruling of the European Commission (EC) concerning pre-installed apps that Google requires to be on mobile devices with the Android 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.

IV-Strategy based on Intensity of Network Effects: An alternative source of identification is variation in the intensity of a sub market's network effects. Some services that apps provide can reach their full service quality independently of whether others use the same

[1] See `http://europa.eu/rapid/press-release_IP-16-1492_en.htm`, last accessed on 12.12.2016.

app or not (e.g. a wallpaper, or apps for producing baby soothing white noise). 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.[2] 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. Examples could be sharing a nice photo or a successful run with ones peers who also use the same app.

In this identification approach we exploit the fact that the strength of a sub market's network effects depend 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 'program' the market's concentration 'in equilibrium,' and are also a good predictor of the observed market concentration, as they converge to their equilibrium state. While this phenomenon creates a fascinating research opportunity in and of itself, the strength of the network effects can thus serve as a technological instrumental variable that exogenously affects concentration. We plan to leverage this opportunity by quantifying the intensity of each submarket's network effects and using this variable in an IV-strategy.

Robustness: We can test the robustness of our results across varying measures of competition, innovation and data collection practices of firms, 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.

[2]Being the single user of a dating app, arguably renders the app somewhat useless.

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: **Innovation:** as measured vie the information on updates:

- date,
- textual information on what is new,
- version number

Competitor information: (for market definition)

- the names and IDs of similar apps

Data collection/Privacy:

- all permissions that apps are requesting (upon installation) and that apps require to perform certain functions (in total approx. 140 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.)

Controls and app-success measures:

- the total number of installations of an app,
- (monthly) downloads of an app,
- 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,

Additional Developer-specific information in the data:

- 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.[3] In our case, we have detailed information on product updates and new product

[3]Web-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.

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 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.

Taken together, our data are superior to standard data sets, because they allow us to study the innovation frequency, innovation content, and innovation quality (rather than only the binary existence of an innovation). Moreover, we can run these analyses for both radical and incremental innovation activity. 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 is the availability of information on app-specific competitor apps. Google's Play

Store provides information on a set of "similar apps" for each app. "Similar 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.[4] We can use the IDs and names of "similar apps" in our data set (which covers the full population of available apps) to identify each app's closest competitors. Using this information about competitors, we can define the set of relevant sub market, and the competitors in the market in two different ways.

App-specific Market Definition: First, we can construct app-specific measures of competition and market power for this limited amount of competitors of up to 24 with the information on the individual download numbers and ratings. This is our first naive measure. This approach corresponds to drawing a circle with a fixed radius around a plant or productive unit.

Clusters of Apps in the Competitor Network: Second, we can construct a network based on the information about "similar apps", that is provided by Google. Each app represents a node in this network and an edge (or link) between two apps is established if one app is listed as a "similar app" of another one. Using network methodology, we are able to detect clusters of apps (or 'isolated app communities') that represent market segments.[5] An exemplary cluster comprising apps related to virtual private networks is displayed in figure 6. Moreover we can exploit additional information (such as affiliation to the same category) in the data to impose efficiency-enhancing restrictions onto the "similar apps" and the resulting clusters. Based on this alternative market defintion, we can compute a second market share for each app, and the resulting market concentration measures.

Being able to construct such detailed 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 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 retrieved from the platform AppBrain.com enabling to study unexpected market dynamics.

Thus, our data set provides an exceptionally rich source for empirically studying the consequences

[4] The actual set of competitor apps is even bigger in several cases.

[5] 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 estimations.

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 Collection by App Developers

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 consider 'actively' collected data (using the permissions an app developer can request upon installation) as a source of information which is potentially available to firms.
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).[6] Grades were assigned using a privacy model that measures the gap

[6] Their results are provided publicly: see: `http://privacygrade.org/`

between people's expectations of an app's behavior and the app's actual behavior.

The information on third-party libraries[7] 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.

4.2 Descriptive Evidence

Table 1: Summary Statistics of Cross-Section and Panel

	Cross Section				Balanced Panel			
	mean	p50	min	max	mean	p50	min	max
HHI	0.22	0.18	0	1	0.20	0.17	0	1
Market Share	0.01	0.00	0	1	0.01	0.01	0	1
$\#_{Updates}$	0.64	0.00	0	24	0.63	0.00	0	20
D_{Update}	0.29	0.00	0	1	0.29	0.00	0	1
$D_{MajorUpdate}$	0.09	0.00	0	1	0.09	0.00	0	1
$D_{DataCollection}$	0.55	1.00	0	1	0.51	1.00	0	1
$\#_{DataCollection}$	1.31	1.00	0	12	1.20	1.00	0	10
D_{Paid}	0.03	0.00	0	1	0.03	0.00	0	1
$\#CleanPerms.$	4.93	4.00	0	177	4.64	4.00	0	42
$D_{InappProduct}$	0.13	0.00	0	1	0.09	0.00	0	1
Avg.Rating	4.06	4.10	1	5	4.15	4.20	1	5
Length Description	1296.12	945.00	4	11588	1358.77	943.00	4	11276
Observations	164093				100305			

Before turning to the main analysis, we provide first descriptive evidence. Table 1 contains summary statistics of the main variables of interest for both the cross section from September 2016 and for the balanced panel. The HHI on the average app is 0.2, but for both cross section and panel, the measure varies from zero to one.[8] A single app's average market share is 1%, but also the market share can take values of one, which corresponds to an app having monopoly power in its segment. In total, there are 9225 clusters with sizes varying from 5 to 1000 apps. The average cluster consists of 55 apps. At the cluster-level the HHI is on average 0.66, ranging from 0.018 to 1. The market concentration is even higher for clusters in the Communication category, whereas it is lower in clusters of the categories Finance or Weather. Innovation is captured by updates, and we both count the number of updates and generated an indicator

[7] 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 `http://privacygrade.org/third_party_libraries` for an exemplary overview.)

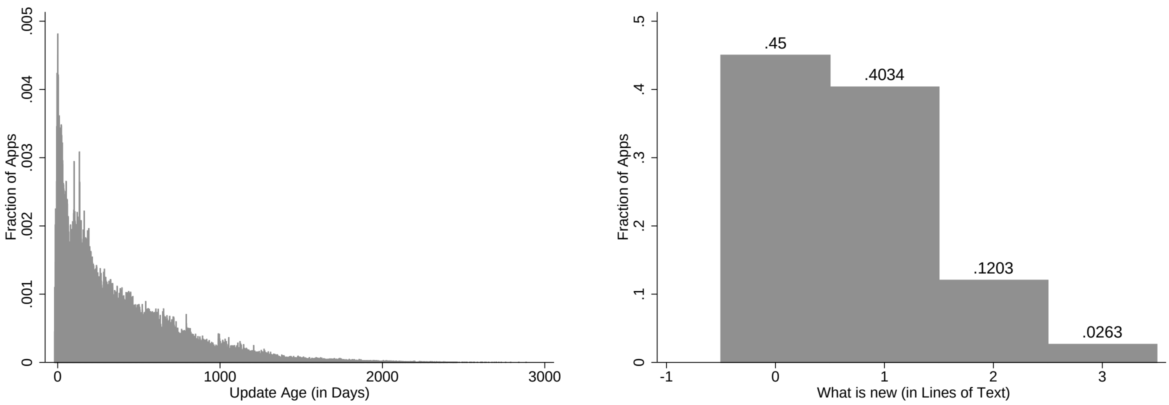
[8] This HHI is calculated based on nearest neighbors (i.e. "similar apps") only.

variable that takes the value of 1 if any updates are observed. We also distinguish regular updates and major updates and generated an additional dummy that indicates if we observe a major update in the given quarter. For both datasets updates occur for approximately 29% of the observations, which means that around a third of the apps provide an update every period. Major updates can be observed for just under 10% of the apps in each period, and this ratio is, again the same in the cross section and the panel data. As far as data collection is concerned, we observe that more than half of the apps collect data. The fact that the average number of sensitive permissions is above one, suggests that conditional on collecting data an app accesses more than two sensitive permissions on average. The table also shows summary statistics for our most important control variables. In-app purchases are less common than one might think, as they are only used by roughly 10% of the apps. Average ratings are skewed around four stars, as it is typical in online ratings. Figures 1, 2 and 3 complete the picture by providing histograms illustrating the distribution of our main innovation, data access and competition variables. Not very surprisingly, data collection and innovation frequency follow a long-tail distribution, but the histogram for the length of the "what is new" section shows nicely how this variable is distributed. The competition measures follow more standard distributions, especially the log. market share.

Figure 1: Histograms of Innovation Variables

(a) Innovation Frequency (Update Age)

(b) Innovation Size (Size of What is New-Section)

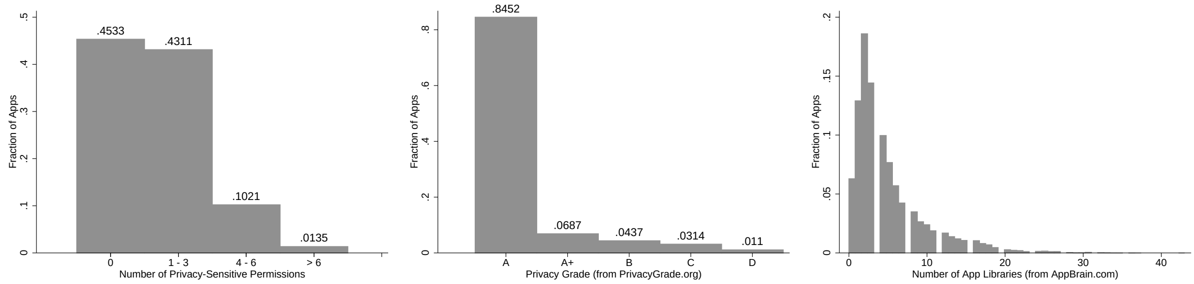


Notes: The figures show, based on the cross-section from September 2016, histograms for two innovation measures. Panel (a) shows the distribution of apps' update age, i.e. the days since the last update. Panel (b) shows the distribution of the sizes of the 'what is new'-section of each app.

In addition, Figures 4 and 5 present the relationship between our market concentration measure, the HHI, and some of the innovation and data access variables. The bar graphs illustrate the mean deviation from the overall average for each HHI quantile and are computed using the cross-sectional data set from September 2016.

Figure 2: Histograms of Data Collection

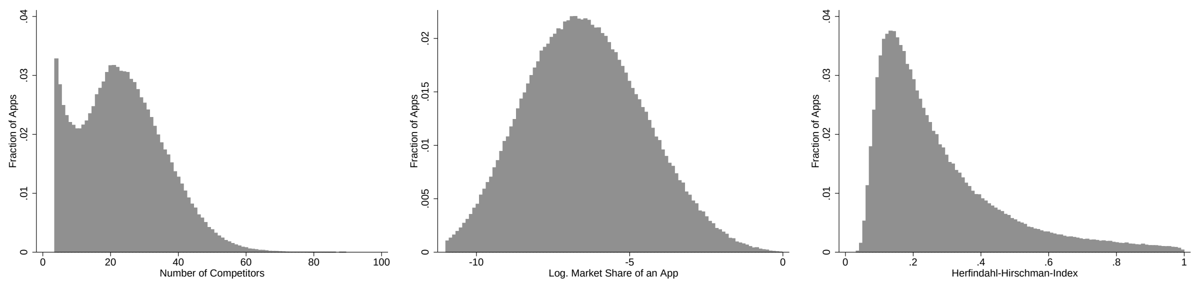
(a) Data Collection (Number of Permissions) (b) Data Collection (Privacy Grade) (c) Data Collection (Number of App Libraries)



Notes: The figures show, based on the cross-section from September 2016, histograms for three measures of data access. Panel (a) shows the distribution of the number of permissions apps request. Panel (b) shows the distribution of the privacy grades (obtained from PrivacyGrade.org), and panel (c) shows the distribution of the number of app libraries an app uses (taken from AppBrain.com).

Figure 3: Histograms of Competition Variables

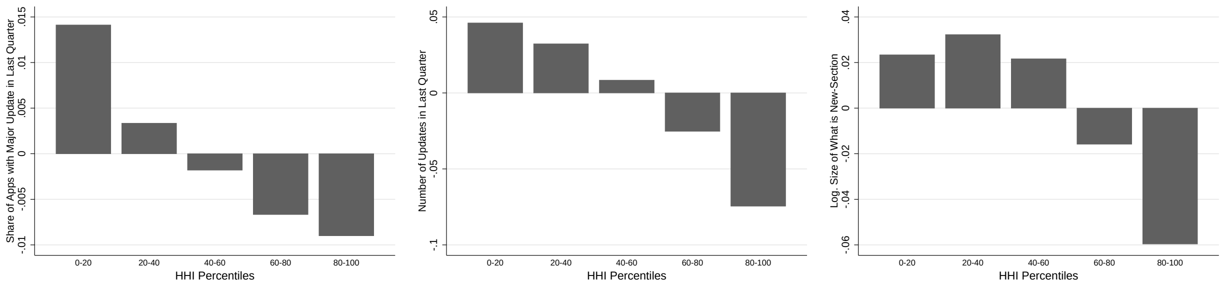
(a) Number of Competitors (b) Log. Market Share (c) Herfindahl-Hirschman-Index



Notes: The figures show, based on the cross-section from September 2016, histograms for three measures of apps' market power and competitive situation. Panel (a) shows the distribution of the number of competitors of an app. Panel (b) shows the distribution of the log market share (computed using the number of additional ratings as a demand measure). Panel (c) shows the distribution of the app-specific Herfindahl-Hirschman-Index.

Figure 4: Bar Graphs of the Relationship between Market Concentration and Innovation

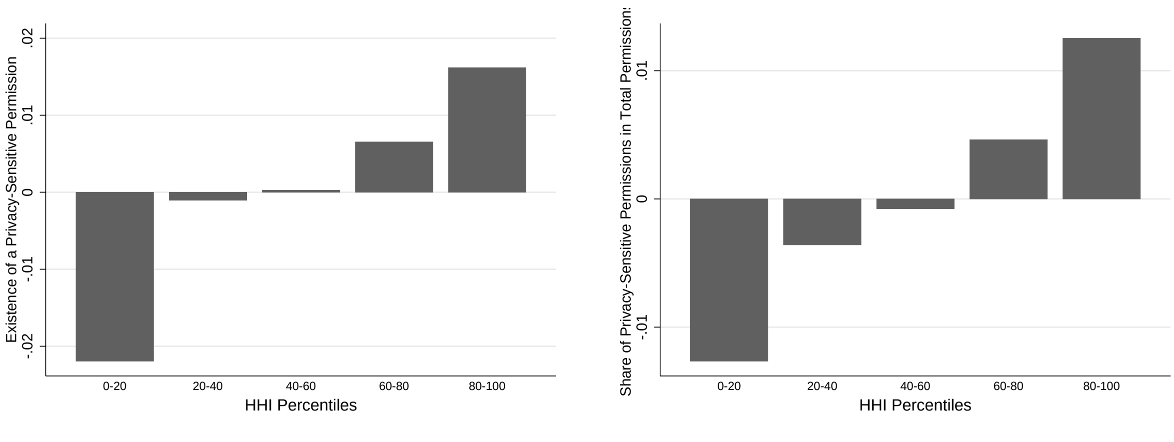
(a) Innovation Frequency (Major Update in Last Quarter) (b) Innovation Frequency (Number of Updates in Last Quarter) (c) Innovation Size (Log. Size of What is New-Section)



Notes: The figures show, based on the cross-section from September 2016, bar graphs for the relationship between HHI and two innovation measures. The figures show the mean deviation of the apps in each HHI-quantile from the overall average value of the respective innovation measure. Panel (a) shows major updates and panel (b) shows update frequency, i.e. with the number of updates of an app in the last quarter. Panel (c) shows the update size, i.e. with the log size of the what is new-section of apps being updated in the last quarter.

Figure 5: Bar Graphs of the Relationship between Market Concentration and Data Access

(a) Data Collection (Existence of Privacy-Sensitive Permissions) (b) Data Collection (Share of Privacy-Sensitive Permissions)



Notes: The figures show, based on the cross-section from September 2016, bar graphs for the relationship between market shares and measures of data access. Both figures show the mean deviation of the apps in each HHI quantile from the overall average value of the respective data measure. Panel (a) shows the relationship with data collection behavior, i.e. a dummy which is equal to one if an app has at least one privacy-sensitive permission, whereas panel (b) uses the share of privacy-sensitive permissions in the total number of permissions.

The link between innovation measures and HHIs is depicted in Figure 4, where panel (a) and (b) show mainly a decreasing trend between innovation frequency, as measured by the share of apps with a major update in the last quarter or the number of updates in that quarter, and the HHI of the apps' market. Regarding the innovation size variable, panel (c) shows a similar overall pattern but with less monotonic behavior across HHI percentiles.

The set of graphs related to data collection is illustrated in Figure 5 which depicts information on data collection. Panel (a) and panel (b) show the mean deviation from the overall average of the existence of privacy-sensitive permissions (indicator variable) and the share of total privacy-sensitive permissions with respect to the total number of permissions requested by the developer, respectively. Overall, there is a positive link between data access and market share, that is, apps with higher market shares tend to request more privacy-sensitive information.

5 Estimation Results

Table 2 shows the conditional correlations between data collection and market concentration. The table shows cross section (in columns 1 & 4) and panel results (columns 2, 3, 5 & 6). The dependent variable is either a dummy equal to one if the app uses at least one permission allowing it to collect information about the app user ($D_{DataCollection}$, in columns 1-3) or is equal to the number of such permissions ($\#_{DataCollection}$, in columns 4-6). Columns 1, 2, 4, & 5 provide results based on the simple market definition weighting all similar apps of an app equally, whereas columns 3 & 6 use our alternative market definition based on the network approach. The results suggest a somewhat positive relationship between market concentration and data collection, especially when using an app-specific market definition. Apps that operate in a more concentrated market seem to collect more data about users. These first results would be in line with the idea of market power resulting in higher data access for app developers. However, the results are considerably weaker when using panel data analysis, and disappear when we base the market definition on clusters in the network of "similar apps."

In Table 3 we analyze the relationship between market concentration and app innovation. The table contains an analysis of the conditional correlations between market concentration and the innovative activity of apps (measured via update behavior). The dependent variable are our measures of innovation: In columns 1-2 we use an indicator equal to 1 if the app provides at least one update in the next wave ($t + 1$), and in columns 3-8 the dummy equals 1 if the app

Table 2: Data Collection and Market Concentration

	$D_{DataCollection}$			$\#_{DataCollection}$		
	CS1	Panel1	Panel2	CS1	Panel1	Panel2
HHI	0.042***	0.008	0.010	0.236***	0.029**	-0.007
	(0.006)	(0.005)	(0.009)	(0.019)	(0.014)	(0.026)
Market Share	-0.035	-0.092*	-0.111***	-0.126	-0.270	-0.409***
	(0.050)	(0.053)	(0.037)	(0.158)	(0.195)	(0.112)
Market Share Squared	0.013	0.152	0.097*	0.269	0.457	0.359**
	(0.131)	(0.146)	(0.057)	(0.396)	(0.623)	(0.182)
D_{Paid}	-0.031***	-0.018	-0.148	-0.057***	-0.022	-0.020
	(0.005)	(0.013)	(0.108)	(0.014)	(0.116)	(0.091)
$\#CleanPerms.$	0.068***	0.059***	0.051***	0.318***	0.266***	0.271***
	(0.001)	(0.002)	(0.002)	(0.005)	(0.006)	(0.007)
$D_{InappProduct}$	0.101***	-0.026*	0.022	0.088***	0.002	0.087***
	(0.003)	(0.013)	(0.014)	(0.009)	(0.028)	(0.034)
Avg.Rating	-0.043***	-0.011*	-0.017**	-0.151***	-0.051***	-0.065***
	(0.002)	(0.006)	(0.007)	(0.005)	(0.019)	(0.022)
Log. Length Description	0.013***	-0.002	0.002	0.003	0.028	0.016
	(0.001)	(0.006)	(0.006)	(0.004)	(0.018)	(0.020)
Constant	0.303***	0.320***	0.380***	0.026	0.045	0.249
	(0.012)	(0.048)	(0.051)	(0.031)	(0.144)	(0.164)
Category	Yes	No	No	Yes	No	No
Wave	No	Yes	Yes	No	Yes	Yes
Observations	239566	162590	150010	239566	162590	150010
Num. of Groups		32518	30002		32518	30002
Adjusted R^2	0.27	0.09	0.07	0.47	0.22	0.22

Notes: This table shows cross-section and panel results for the relationship between data collection and market concentration. The dependent variable is either a dummy equal to one if the app uses at least one permission allowing it to collect information about the app user ($D_{DataCollection}$) or is equal to the number of such permissions ($\#_{DataCollection}$). The variable of interest is a HHI, a measure of market concentration. Heteroscedasticity-consistent standard errors in parentheses: *** p<0.01, ** p<0.05, * p<0.1.

provides at least one major update (also in $t+1$). The main variable of interest the HHI, our measure of market concentration. To provide first evidence whether these additional user data allow app developers to more and larger updates, we also add our measures of data collection ($D_{DataCollection}$ or $\#_{DataCollection}$). In all specifications we control for a squared polynomial function of an app's market share. The results indicate a strong and robust negative relationship between market concentration and innovation. Moreover, data access and innovation seem to be in a robust positive relationship.

Table 3: Innovation, Data Access and Market Concentration

	$Update_{t+1}$		$MajorUpdate_{t+1}$					
	CS1	Panel1	CS1	CS1	Panel1	Panel1	Panel2	Panel2
HHI	-0.103***	0.015	-0.036***	-0.035***	-0.019*	-0.019*	-0.058***	-0.058***
	(0.005)	(0.015)	(0.003)	(0.003)	(0.011)	(0.011)	(0.017)	(0.017)
Market Share	1.907***	-0.021	0.614***	0.614***	-0.384***	-0.384***	-0.560***	-0.569***
	(0.057)	(0.132)	(0.037)	(0.037)	(0.122)	(0.124)	(0.064)	(0.064)
Market Share Squared	-3.217***	-0.229	-1.012***	-1.011***	0.456	0.448	0.578***	0.585***
	(0.182)	(0.298)	(0.117)	(0.116)	(0.342)	(0.353)	(0.097)	(0.096)
$\#_{DataCollection}$	0.005***	0.004	0.005***		0.059***		0.059***	
	(0.001)	(0.005)	(0.000)		(0.005)		(0.005)	
$D_{DataCollection}$				0.011***		0.120***		0.106***
				(0.001)		(0.013)		(0.013)
D_{Paid}	0.032***	0.151	0.003	0.003	0.216**	0.221**	-0.022	-0.011
	(0.005)	(0.103)	(0.003)	(0.003)	(0.102)	(0.106)	(0.053)	(0.057)
$\#CleanPerms.$	0.015***	-0.015***	0.004***	0.005***	-0.067***	-0.058***	-0.069***	-0.059***
	(0.001)	(0.003)	(0.000)	(0.000)	(0.003)	(0.003)	(0.003)	(0.003)
$D_{InappProduct}$	0.046***	-0.057***	0.008***	0.007***	-0.115***	-0.112***	-0.106***	-0.106***
	(0.003)	(0.021)	(0.002)	(0.002)	(0.021)	(0.021)	(0.019)	(0.019)
Avg.Rating	0.006***	-0.026*	0.001	0.001	0.019	0.018	-0.016	-0.017
	(0.001)	(0.015)	(0.001)	(0.001)	(0.013)	(0.013)	(0.012)	(0.012)
Log. Length Description	0.025***	-0.047***	0.006***	0.005***	-0.200***	-0.198***	-0.225***	-0.224***
	(0.001)	(0.011)	(0.000)	(0.000)	(0.011)	(0.011)	(0.012)	(0.012)
Constant	-0.249***	0.818***	-0.052***	-0.056***	1.632***	1.589***	2.006***	1.972***
	(0.008)	(0.097)	(0.005)	(0.005)	(0.095)	(0.096)	(0.096)	(0.096)
Category	Yes	No	Yes	Yes	No	No	No	No
Wave	No	Yes	No	No	Yes	Yes	Yes	Yes
Observations	225274	130072	225274	225274	130072	130072	120008	120008
Num. of Groups		32518			32518	32518	30002	30002
Adjusted R^2	0.08	0.00	0.02	0.02	0.03	0.03	0.08	0.07

Notes: This table shows cross-section and panel results for the relationship between market concentration, data collection and innovation. The dependent variable is a measure of innovation: Either it is a dummy equal to one if the app provides at least one update in the following wave ($t+1$) or if it provides at least one major update in the following wave. The variables of interest are: (1) a HHI, a measure of market concentration, and (2) either a dummy which is equal to one if the app has at least one permission allowing it to collect information about the app user ($D_{DataCollection}$) or the number of such permissions ($\#_{DataCollection}$). Heteroscedasticity-consistent standard errors in parentheses: *** p<0.01, ** p<0.05, * p<0.1.

Beyond the analysis of the conditional correlations we plan to us identification strategies based on exogenous variation in the Play Store. Specifically we plan to use Google's promotional activity (named "Deal of the Week"), and unpredictable market entry of new bestseller apps, as exogenous shifters of market structure. Similarly, we want to exploit natural variation in the strength of a sub market's network effects in an instrumental variables strategy. We are currently implementing these rigorous approaches to identifying the causal effect of market structure on innovation and data collection. Finally, in Table 4 we provide robustness checks for the relationship between data collection, market concentration and innovation. We use logs to normalize the distributions of market share and the HHI, and we split our sample into six subgroups: top apps, average apps, free and paid apps as well as non-game and game apps. Overall, these results support our baseline findings: data collection comes with higher subsequent innovation success, whereas a more concentrated market results in lower innovation activity.

Discussion, Limitations and Further Research

Our findings above are only first results, and they highlight only the conditional correlations between market concentration and innovation in the market for mobile applications. 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.

Altogether, we clearly confirm the hypotheses that market concentration and innovation are closely related. We also see that data appears to play an important role for innovation. However, while our evidence is suggestive of a direct relationship between market concentration and data collection, this relationship is somewhat less robust than initially expected, and could also be driven by market power.

Further research will attempt to further uncover the causal relationship by using exogenous variation. Our primary strategy aims at exploiting Google's promotional activity (named "Deal of the Week"), and unpredictable market entry of new bestseller apps, the popularity of which is hard to predict for incumbents (and for the developers of the new apps alike). A second exogenous shock could soon arise from 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. Moreover, such shocks will allow us to investigate more deeply whether it is concentration or rather market power that drives the relationship between market power and data access.

Alternatively, we can use the underlying 'natural strength' of a market segment's network effects to devise an instrumental variable that exogenously drives market concentration. 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 might have more access to user data through active data collection, which might imply less user privacy. That access to user data might, in turn, 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 more access to data via data collection. At the same time, apps which have better access to user data also show a higher innovation activity. They are more likely to be updated and their updates are larger. 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 is based on exogenous variation. We will exploit information on unexpected market entries of rivals, use information on promotional activities by Google, and attempt to leverage policy changes implemented by Google. In parallel we will implement an IV-strategy that leverages natural variation in the strength of the network effects in the sub markets.

References

Aghion, P., N. Bloom, R. Blundell, R. Griffith, and P. Howitt, "Competition and Innovation: An Inverted-U Relationship," *Quarterly Journal of Economics*, 2005, *120* (2), 701–728.

AppAnnie, "AppAnnie Mobile App Forecast: The Path to $ 100 Billion," Technical Report 2016.

Arora, A., C. Forman, A. Nandkumar, and R. Telang, "Competition and Patching of Security Vulnerabilities: An Empirical Analysis," *Information Economics and Policy*, 2010, *22* (2), 164–177.

Arrow, K., "Economic Welfare and the Allocation of Resources for Invention," in "The Rate and Direction of Inventive Activity: Economic and Social Factors" NBER Chapters, National Bureau of Economic Research,, 1962, pp. 609–626.

Askalidis, G., "The Impact of Large Scale Promotions on the Sales and Ratings of Mobile Apps: Evidence from Apple's App Store," Technical Report 2015.

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

Boudreau, K. J., "Let a Thousand Flowers Bloom? An Early Look at Large Numbers of Software App Developers and Patterns of Innovation," *Organization Science*, 2012, *23* (5), 1409–1427.

_ , L. B. Jeppesen, and M. Miric, "Competition and Entry in Digital Markets," Technical Report, Mimeo 2015.

Boudreau, K.J. and L.B. Jeppesen, "Unpaid Crowd Complementors: The Platform Network Effect Mirage," *Strategic Management Journal*, 2015, *36* (12), 1761–1777.

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

Brynjolfsson, E. and K. McElheran, "Data in Action: Data-Driven Decision Making in U.S. Manufacturing," Technical Report, US Census Bureau Center for Economic Studies Paper No. CES-WP-16-06 2016.

_ **, L. M. Hitt, and H. H. Kim**, "Strength in Numbers: How Does Data-Driven Decision Making Affect Firm Performance?," Technical Report 2011.

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

Chaudhari, H., "The Impact of Zero-Price Promotions on Sales: A Case Study of Amazon Appstore," Technical Report 2015.

Cohen, W. 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.

_ **and R. C. Levin**, "Empirical Studies of Innovation and Market Structure," in R. Schmalensee and R. Willig, eds., *Handbook of Industrial Organization*, Vol. 2 of *Handbook of Industrial Organization*, Elsevier, 1989, pp. 1059–1107.

Crepon, B., E. Duguet, and J. Mairessec, "Research, Innovation And Productivity: An Econometric Analysis at the Firm Level," *Economics of Innovation and New Technology*, 1998, *7* (2), 115–158.

Dasgupta, P. and J. Stiglitz, "Industrial Structure and the Nature of Innovative Activity," *The Economic Journal*, 1980, *90* (358), 266–93.

Foerderer, J., T. Kude, S. Mithas, and A. Heinzl, "Does Platform Owner's Entry Crowd Out Innovation? Evidence from Google Photos," Technical Report 2016.

Jo, A. M., "The Effect of Competition Intensity on Cybersecurity - An Empirical Analysis of Security Patch Release on the Web Browser Market," Technical Report 2016.

Kummer, M. E. and P. Schulte, "When Private Information Settles the Bill: Money and Privacy in Google's Market for Smartphone Applications," ZEW Discussion Papers 16-031, ZEW / Center for European Economic Research 2016.

Lin, J., B. Liu, N. Sadeh, and J. I. Hong, "Modeling Users Mobile App Privacy Preferences: Restoring Usability in a Sea of Permission Settings," in "Symposium On Usable Privacy and Security 2014" 2014, pp. 199–212.

_ , S. Amini, J. I. Hong, N. Sadeh, J. Lindqvist, and J. Zhang, "Expectation and Purpose: Understanding Users' Mental Models of Mobile App Privacy Through Crowdsourcing," in "Proceedings of the 2012 ACM Conference on Ubiquitous Computing" ACM 2012, pp. 501–510.

Liu, Y., D. Nekipelov, and M. Park, "Timely Versus Quality Innovation: The Case of Mobile Applications on iTunes and Google Play," Technical Report, NBER Working Paper 2014.

Miric, M., "Essays on Competition, Innovation and Firm Strategy in Digital Markets." PhD School of Economics and Management, Copenhagen Business School 2015.

Mohnen, P., J. Mairesse, and J. M. Dagenais, "Innovativity: A Comparison Across Seven European Countries," *Economics of Innovation and New Technologies*, 2007, *15* (4-5), 391–413.

OECD and Eurostat, *Oslo Manual - Guidelines for Collecting and Interpreting Innovation Data* OECD and Eurostat, 3rd ed. 2005.

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

Sarma, B. P., N. Li, C. Gates, R. Potharaju, C. Nita-Rotaru, and I. Molloy, "Android Permissions: A Perspective Combining Risks and Benefits," in "Proceedings of the 17th ACM symposium on Access Control Models and Technologies" ACM 2012, pp. 13–22.

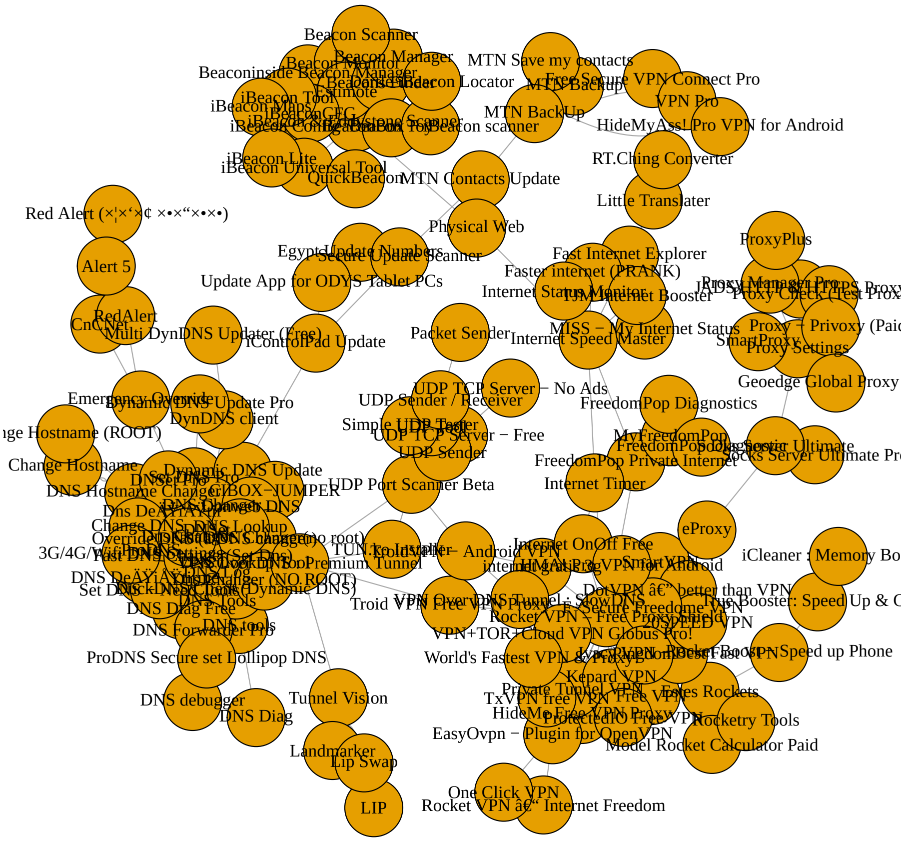
Saunders, A. and P. Tambe, "Data Assets and Industry Competition: Evidence from 10-K Filings," Technical Report 2015.

Tambe, P., "Big Data Investment, Skills, and Firm Value," *Management Science*, 2014, *60* (6), 1452–1469.

7 Appendix

Data

Figure 6: Exemplary Cluster of Similar Apps



Notes: The figure shows the cluster of apps that allow users to access the internet through a vpn-access. 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 network cluster or sub market in our analysis, some apps are clearly closer in the "similar app" network than others.

Robustness

Table 4: Innovation, Data Access and Market Concentration

	Top&Avg.Apps		Free&Paid		Normal&Game	
	(1)	(2)	(3)	(4)	(5)	(6)
Log. HHI	-0.011**	-0.008***	-0.010***	-0.018	-0.010***	-0.001
	(0.005)	(0.003)	(0.002)	(0.014)	(0.003)	(0.011)
Log. Market Share	-0.011***	-0.016***	-0.017***	-0.018**	-0.018***	-0.008
	(0.003)	(0.002)	(0.002)	(0.008)	(0.002)	(0.007)
$\#_{DataCollection}$	0.064***	0.061***	0.061***	0.012	0.060***	0.107***
	(0.007)	(0.007)	(0.005)	(0.031)	(0.005)	(0.030)
D_{Paid}	0.308***	0.193*	0.000	0.000	0.196*	0.000
	(0.026)	(0.113)	(.)	(.)	(0.105)	(.)
$\#CleanPerms.$	-0.065***	-0.077***	-0.070***	-0.052***	-0.070***	-0.116***
	(0.005)	(0.004)	(0.003)	(0.016)	(0.003)	(0.020)
$D_{InappProduct}$	-0.164***	-0.103***	-0.152***	0.049	-0.147***	-0.083
	(0.025)	(0.031)	(0.020)	(0.091)	(0.021)	(0.069)
Avg.Rating	0.031	0.001	-0.000	-0.073	-0.001	0.031
	(0.036)	(0.012)	(0.011)	(0.068)	(0.012)	(0.058)
Log. Length Description	-0.209***	-0.197***	-0.205***	-0.237***	-0.209***	-0.341***
	(0.016)	(0.016)	(0.011)	(0.079)	(0.012)	(0.060)
Constant	1.663***	1.567***	1.656***	2.278***	1.683***	2.669***
	(0.187)	(0.119)	(0.090)	(0.657)	(0.099)	(0.471)
Wave	Yes	Yes	Yes	Yes	Yes	Yes
Observations	44444	85337	125950	4122	117837	12235
Num. of Groups	14668	24338	31492	1037	31920	8821
Adjusted R^2	0.09	0.07	0.08	0.07	0.07	0.09

Notes: This table shows cross-section and panel results for the relationship between market concentration, data collection and innovation. The dependent variable is a measure of innovation: Either it is a dummy equal to one if the app provides at least one update in the following wave ($t + 1$) or if it provides at least one major update in the following wave. The variables of interest are: (1) a HHI, a measure of market concentration, and (2) the number of permissions allowing an app to collect information about the user ($\#_{DataCollection}$). In this table, we do split the sample into various subsamples: (1) top and average apps, where top apps are those with more than 100 additional ratings in a wave, and average apps are the remaining ones; (2) free and paid apps; (3) non-game and game apps. Heteroscedasticity-consistent standard errors in parentheses: *** p<0.01, ** p<0.05, * p<0.1.