10.5465/AMBPP.2020.203

COMPETITION AND PRIVACY IN ONLINE MARKETS: EVIDENCE FROM THE MOBILE APP INDUSTRY

REINHOLD KESLER
University of Zurich
Plattenstrasse 14
CH-8032 Zurich, Switzerland

MICHAEL E. KUMMER
University of East Anglia

PATRICK SCHULTE
Deutsche Bundesbank

ABSTRACT

Several major countries are planning profound regulatory changes to address growing concerns related to market power and data collection in digital markets. We construct a panel data set that allows us to observe data collection by more than 1.5 million mobile applications over two years. We use a novel network-based approach to define several thousand submarkets on Google’s Play Store and characterize apps’ competitive environments. We are the first to document a robust positive relationship between market concentration and data collection. Our results inform the ongoing policy debate about privacy and competition between antitrust authorities, policymakers, and the public.

MOTIVATION

The dominance of some products and firms in digital markets today often comes along with a massive and systematic collection of personal information. This development raises concerns by numerous policy makers as consumer harm might arise from market power enabling exploitative extraction of consumer data. Decision makers have been particularly alarmed by app developers’ collection and use of personal user data, and the fact that some of these firms have considerable market power. As a result, several major countries are considering to redefine their legal and regulatory frameworks for the digital economy.

Although a growing number of theoretical studies suggest that more market power allows companies to collect more data, empirical studies of the relationship between competition and privacy are difficult. Hence, the evidence to date is based on small-scale samples and correlational approaches (Preibusch & Bonneau, 2013). The present paper provides first large-scale empirical evidence for one of the most important markets with respect to data collection, the mobile app industry. Furthermore, we attempt to expand on the existing research, by identifying causal effects.

RELATED LITERATURE

The paper contributes to the theoretically well-established research on data collection (or the intrusion of individuals' privacy) in the presence of competition or lack thereof (Campbell,

10.5465/AMBPP.2020.203

Goldfarb, & Tucker, 2015; Casadesus-Masanell & Hervas-Drane, 2015; Dimakopoulos & Sudaric, 2018). Furthermore, it can be embedded into a larger context on the provision of product quality in the absence of competition, as a greater intrusion of privacy reduces the quality of a product.

It is also related to the research on the value of privacy (Acquisti, Taylor, & Wagman, 2016; Brown, 2016). More specifically, respective empirical studies examine the effects of protecting individuals' privacy through the use of cookies (Johnson, Shriver, & Du, 2020; Marotta, Abhishek, & Acquisti, 2019) or privacy policies (Goldfarb & Tucker, 2011; Tucker, 2014; Goldberg, Johnson, & Shriver, 2019). In the context of apps, research attempted to measure the value of privacy on the basis of experiments or observational data (Savage & Waldman, 2015; Kummer & Schulte, 2019). However, none of the previous studies took into account the impact of the competitive situation and market structure, which is of particular importance in this paper. Accordingly, the relationship between competition (or the lack of competition and thus concentrated market power) on the one hand and the collection of sensitive user data in the online market for mobile apps on the other hand will be analyzed. In this way, the paper contributes by providing first large-scale empirical evidence in a relevant market with novel measures and attempts to identify causal effects.

DATA

For the analysis, almost all apps in the Google Play Store were web-scraped quarterly from October 2015 to January 2018, thereby giving us all the publicly available information about each app and its developer.

A special challenge in the case of online markets is the definition and identification of markets and the competitors operating in them. A starting point for defining these may be provided by the categories of the Google Play Store, similar to studies at the industry level. Google distinguishes close to fifty categories, but these are very broad in scope, sometimes containing more than a hundred thousand apps that do not all compete directly with each other. For this reason, available information about similar apps, which are displayed for each app, is better suited to identify relevant sub-markets. The selection of similar apps is made by Google and leads to the identification of up to fifty comparable products; the actual number of competing apps may be greater. Based on this information, a network is formed that reflects the entire app market. Each app represents a node in the network; a connection between two apps is established when one app is listed as a similar app of another. Using network analyses, we are able to identify several thousand sub-markets as isolated clusters of apps with a common theme. Our approach was adopted by other studies (Wen & Zhu, 2019; Ershov, 2018a). Having defined app-specific markets, apps' market shares and measures of market concentration are calculated based on apps' number of ratings and installations which can be considered as proxies of demand.

The extent to which apps have access to user data is also difficult to observe. However, the Google Play Store – unlike comparable platforms – provides information on apps' permissions to access user data. Accordingly, the paper makes use of these permissions, which are requested by an app and must be accepted before installation by the users. The Google Play Store differentiates between more than 200 permissions, however, our analysis is limited to those that allow for the tracking of user behavior and preferences and can therefore be considered as privacy-sensitive. Based on different classifications, 25 relevant permissions are identified (e.g.

10.5465/AMBPP.2020.203

access to the contacts or the location of the user). Finally, for a subset of apps, we retrieved information on whether data is passed on to third parties via third-party libraries. Although these libraries generally facilitate access to services and help to increase functionality by integrating them into the app, they often require specific user data in exchange, which is why they are classified as intrusive in this context.

In the end, this results in a quarterly panel dataset of around 1.5 million apps, which represent the units of observation and thus map the product level. In addition to measures of data collection and competition, the dataset contains a variety of app and developer characteristics that enable, for example, the measurement of monetization, quality, and functionality.

DESCRIPTIVE EVIDENCE

Summary statistics show that 52 percent of apps have access to user data, while the average app requests 1.33 privacy-sensitive permissions and is active in a market with an HHI equal to 0.15 (averaged across clusters). Additionally, within markets, apps have very different market shares: the average app has a market share of two percent, whereas the median app has a market share of below one percent indicating the long tail towards higher market shares. Other characteristics show that only 7 percent are paid apps and that apps have an average rating of 4.03 out of a possible 5 stars.

The competitive situation in the app market can be illustrated by the number of competitors and the Herfindahl-Hirschmann Index (HHI). The heterogeneity in the identified markets is immense. While 15 percent of the markets consist of 11 to 25 apps, eight percent of all markets have more than 100 apps. A complementing perspective is provided by the HHI as a measure of market concentration, which is calculated on the basis of the number of ratings (or predicted installations). The higher the index is, the more concentrated the market under consideration is. The majority of markets are not very concentrated. Nevertheless, an HHI of over 0.5 occurs for more than 30 percent of the markets considered, which is mainly driven by smaller markets consisting of less than 10 apps. However, one third of this highly concentrated group consists of more than ten apps.

The access and the extent to which privacy-sensitive permissions are requested are the key variables for data collection. One in four apps, have accesses to the location of the users, while more than five percent can view users' contacts. 48 percent of apps do not request permissions that affect users' privacy. One third of the apps access one or two permissions of this type. The proportion of apps requesting more privacy-sensitive permission drops sharply as the number increases, so that only 1.56 percent of the apps require more than six of such permissions.

Looking at the relationship between competition and data collection, our descriptive evidence indicates a positive and monotonous relationship between data collection and market concentration as well as market shares. Put differently, apps in concentrated markets as well as those with a higher market share retrieve on average more user data and thus intrude the privacy of their users more likely. This is not the case when looking at the relationship between competition and unproblematic permissions.

ECONOMETRIC ANALYSES AND RESULTS

10.5465/AMBPP.2020.203

In the first part of our econometric analysis, we confirm the correlational descriptive results in a regression framework. The dependent variable in this framework is the number of requested privacy-sensitive permissions, which is the main factor for data collection. The key independent variables are the measures of market concentration and of apps' market share. Other app characteristics and of their developers are also included in order to avoid biases from differences in monetization strategies, quality, or functionalities. Furthermore, app and time fixed effects are included to take into account the time-constant, unobserved heterogeneity between apps.

The results indicate a positive relationship between data collection and market power. This finding is confirmed in both cross-sectional as well as panel estimates and is independent of the demand measures used. The finding suggests that apps which operate in more concentrated markets and with a higher market share request more data from their users than apps in less concentrated markets or apps with smaller market shares. The other control variables show effects in line with the expectations. Apps that are free of charge or have a high level of functionality are more likely to request more privacy-sensitive permissions.

The relationship is robust to various checks, which can be divided into three groups. Firstly, we vary the measures of data collection and market structure. Taking into account the fact that some permissions are of central importance for the functionality of the app (e.g. location data for the functionality of navigation apps) does not lead to different results.

Secondly, we test the sensitivity of underlying assumptions by limiting the sample to a balanced panel and using the current rather than cumulative demand. Varying the market definition – we use narrow (up to 50 similar apps) or broad categories – does not alter the results qualitatively, either.

Finally, we repeat our analyses, but use data sharing with third parties as a dependent variable. Again, there is a positive correlation with market concentration. This additional result suggests that data in this market seems to be exchanged for monetary benefits. Indeed, the baseline result can only be found in the case of free apps, but not in the case of paid apps, and it is stronger in economically more relevant markets (measured by the total number of installations or apps in that market).

The methods and specifications presented so far do not yet allow for causal statements. More access to data and thus a larger data collection could, for example, lead to an improved market position, which raises reverse causality concerns. Mitigating this problem is the goal of the second part of our econometric analysis, in which we exploit sources of exogenous variation in the competitive environment.

At the heart of the second set of results is the analysis of a redesign of the categories (recategorization), which is used as a natural experiment. Due to the very diverse apps within the categories in the Play Store, Google occasionally adds new categories. Some of the existing apps that belong to the same topic are moved to those new categories. Similar to Ershov (2018b), we exploit the announcement of eight new app categories by Google in July 2016 as an exogenous variation. This procedure is based on the hypothesis that for new categories competitors are more exposed, the costs for search by users decrease, and new players are attracted. These changes would suggest increased competition in those new categories. Due to a lack of detailed information on the part of Google regarding the apps concerned, it was not possible for app developers to anticipate the recategorization.

10.5465/AMBPP.2020.203

Descriptive analyses of the average market concentration and data access for the group of newly assigned apps compared to the remaining apps show the average concentration in markets affected by the category change to decrease. Similarly, subsequent to the recategorization there is a sharp decrease in the number of requested privacy-sensitive permissions in new categories. Further regression analyses exploiting the recategorization also indicate that recategorized apps request less user data after the intervention.

CONCLUSION

The present paper establishes new ways of measuring competition and data collection in the online market for mobile apps. Descriptive results show that a significant share of apps find themselves in highly concentrated markets and that the access to user data is a common phenomenon in this market. The results indicate that apps in more concentrated markets and with a higher market share request more data about users. The robustness of this result is tested in various alternative specifications in which various approaches to measuring the variables are considered. Thus, the paper provides the first large-scale empirical evidence on the relationship between market structure and data collection and gives important input for the current political debate on the role of data, which is being given increasing attention by competition authorities in particular.

REFERENCES ARE AVAILABLE FROM AUTHORS