ISSN 2281-4299

Dipartimento di Ingegneria informatica
automatica e gestionale Antonio Ruberti

Sapienza
Università di Roma

Towards a price for private information of mobile users: the Arcade apps in Google Play Store case

Alessandro De Carolis
Andrea Vitaletti

Technical Report n. 9, 2015

Towards a price for private information of mobile users: the Arcade apps in Google Play Store case.

Alessandro De Carolis [*1] and Andrea Vitaletti[†1]

[1]DIAG, University of Rome "La Sapienza"

September 2, 2015

Abstract

In this paper we present a first attempt to provide an economic value to mobile users' private information. We claim that when users grant access to an application's required permissions, they disclose private information and data to third parties and let them possibly make revenues out of it. To put a monetary value to such information, we use the price of non-free applications. We use a linear model trained on the 5,187 non-free Arcade apps in Google Play Store that takes a set of permissions in input and estimates the corresponding price. Under the assumption that users "pay" free applications by providing access to more private information (i.e. permissions) and consequently the more permissions are required the less users pay the application, our research aim at showing that the estimated price provides a good proxy to attribute a quantitative value to private and sensitive information of mobile apps' users.

Keywords: Privacy, economic value of private information, application's required permissions, machine learning, Google Play Store dataset.

[*] alessandrodc88@gmail.com

[†] vitaletti@dis.uniroma1.it

1 Introduction

The market of mobile applications is continuously expanding. According to the market report “Mobile Applications Market - Global Industry Analysis, Size, Share, Growth, Trends and Forecast, 2014 - 2020”, published by Transparency Market Research, the market was valued at US$ 16.97 Bn in 2013 and is expected to reach US$ 54.89 Bn by 2020, growing at a CAGR of 16.2% from 2014 to 2020. However, according to the 2015 estimates by the independent market research firm eMarketer, only 33% of the U.S. mobile users is available to pay for their mobile apps.

On the basis of this observation a natural question arises: why should developers put their apps on the market if only 1/3 will provide some revenue? There is a number of answers to this question, but we claim that the main reason is that even free applications are paid by users by allowing the disclosure of private information. This allows app providers to make direct or indirect revenues out of that sensible and private information.

In this paper we support this claim with the first quantitative analysis that we carried out on a significant number of applications in the Google Play marketplace. In particular, we show how we can estimate an application’s price by simply observing the public available information on permissions required by that application during the installation step.

The paper is organised as follows: in section 2 we discuss the related works. It is worth noticing, that while there are a number of papers investigating the implications in terms of security due to a misuse of permission in mobile applications, to the best of our knowledge there are no papers trying to estimate the economic value of mobile users’ private information on the basis of the required permissions. In section 3 we introduce our dataset obtained collecting information on the Google Play Store and then in section 4 we use this dataset to obtain a linear model to predict the price of an application on the basis of the required permissions. In section 5 we extend our analysis considering the ranking and finally in section 6 we discuss the obtained results and the future works.

2 Related Works

The paper by Kleinberg [7] introduces the principle underlying privacy that inspires our work: *“Individuals are entitled to control the dissemination of private information, disclosing it as part of a transaction only when they are fairly compensated”*. The paper also shows how this principle can be made precise in several diverse settings in the framework of coalitional games.

In [9] the authors investigated users privacy issues in Android Ad libraries. They observe that mobile applications may have a direct purchasing cost or be free but ad-supported and they examined the effect on user privacy of 13 Android Ad providers. They discovered that several ad libraries can worryingly access permissions beyond the required ones. Similarly, two parties with conflicting interests: the user, interested in maintaining their privacy and the developer who would like to maximize their advertisement revenue through user profiling, are considered in [8].The analysis of more than 250,000 applications in the Android market shows that the current privacy protection mechanisms are not effective. In [5] the authors studied Android applications to determine whether Android developers follow least privilege with their permission requests. Their Stowaway tool, applied to 940 apps, showed that about one-third of Androids apps are overprivileged.

In [10] the authors observe that privacy requirements in mobile apps have to face specific challenges, they are *"highly dynamic, changing over time and locations, and across the different roles of agents involved and the kinds of information that may be disclosed"* and this make often difficult to evaluate the effect of privacy requirements a priori. On the basis of similar considerations, the authors of [4] proposed MockDroid, a modified version of the Android operating system which allows users to revoke access to particular resources at run-time, encouraging users to consider the trade-off between functionality and the disclosure of personal information whilst they use an application.

3 The dataset

We collected the dataset on the Google Play Store, using an ad-hoc application that employs Selenium [3] to mimic the behaviour of a user that wants to install a mobile app available on the store. Selenium is a tool to automate browser actions; it takes in input the url of a page in the store and returns an object that represents that page. The page is parsed to extract useful information (see table 1) that are stored in the dataset. Among such information, the url of similar applications and the url of the applications by the same developer are inserted in the list of the next pages to be analysed. The crawler selects the next page from the list and this process iterates until all the applications have been analysed.

The dataset contains the information summarised in table 1 on 1,135,700 applications.

A detailed explanation of the permissions in Android is available at [1]. Our dataset contains 266 permissions from 15 categories. As an example, the Fruit Ninja application shown in figure 2 requires the permissions: In-app purchases, find accounts on the device from the category identity,modify or delete the contents of your USB storage and read the contents of your USB storage from the category Photos/Media/Files and a number of other permissions from the category Other.

We will focus on the 35,268 applications in the Arcade category. Among

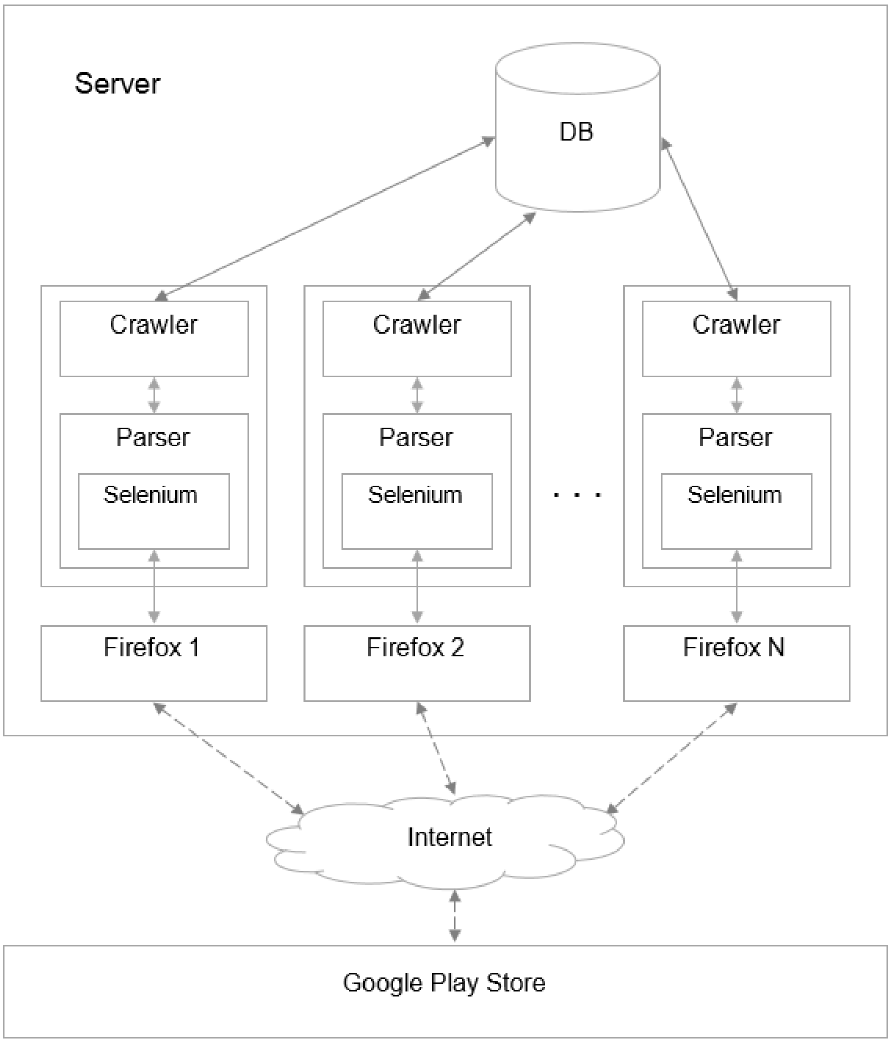
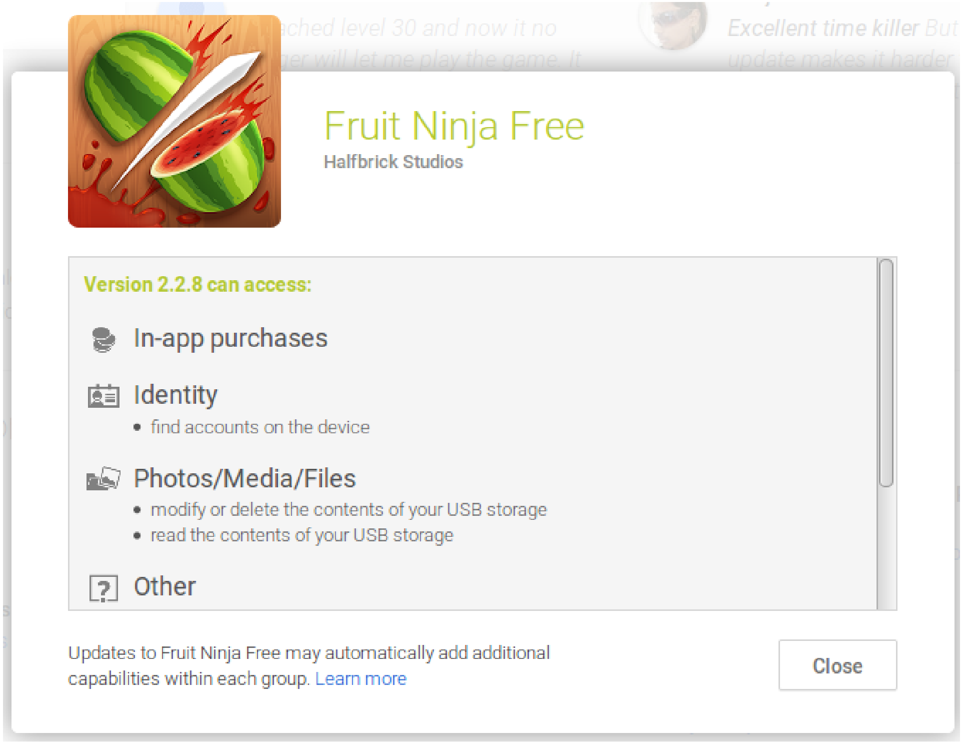


Figure 1: A simplified architecture of the application to collect the data from Google Play web site.



Fruit Ninja Free

Halfbrick Studios

Version 2.2.8 can access:

In-app purchases

Identity

- find accounts on the device

Photos/Media/Files

- modify or delete the contents of your USB storage
- read the contents of your USB storage

Other

Updates to Fruit Ninja Free may automatically add additional capabilities within each group. Learn more

Close
ched level 30 and now it no
ger will let me play the game. It
Excellent time killer But
update makes it harder

Figure 2: An example of the permission required by the well known Fruit Ninja Arcade app.

Table 1: A summary of the main attributes for each application in the dataset.

Name	Name of the application
url	Url of the application main page
developer url	Url of the developer
category	Category of the application
price	The price of the application
inAppPurchaseMin	The minimum expense for a purchase made by the app.
inAppPurchaseMax	The maximum expense for a purchase made by the app.
rating	The average value of the rating provided by the users
number of ratings	The number of ratings for the app
downloads	Number of downloads for the app
permissions	The list of permissions required by the app

these, more than 85% are free and only 5,187 are non-free applications with a price ranging from 0.5 to 137.69 euros. Furthermore, only 1% of the non-free applications have a price > 5 euros. For this reason, in the remaining of the paper we consider such applications as outliers and we focus on the dataset made of non-free applications in the Arcade category with a price ranging between 0.5 and 5 euros. The distribution of prices for such applications is shown in figure 3; most of the applications (about 60%) have a price between 0.5 and 1 euro and more than 80% of the applications cost less than 2 euros.

Figure 4 shows the 30 most used frequency in the considered dataset. More than 60% of the applications use the *full network access* permission. All the other permissions are used in less than 50% of the application.

4 Estimating the price of private information

In this section we discuss our first attempt to estimate the price of the private information of mobile users.

4.1 The method

Our analysis is based on the method briefly discussed in this section. We claim that when a user grants the access to the permissions requested by an application, she allows third parties to access her private information and possibly make revenues on this. In other words, we formulate the following hypothesis:

H1 The revenues for application providers are given by a mix of a) the price of the application in the store, b) the possibility of accessing users' private information. As much the price of the app is low as

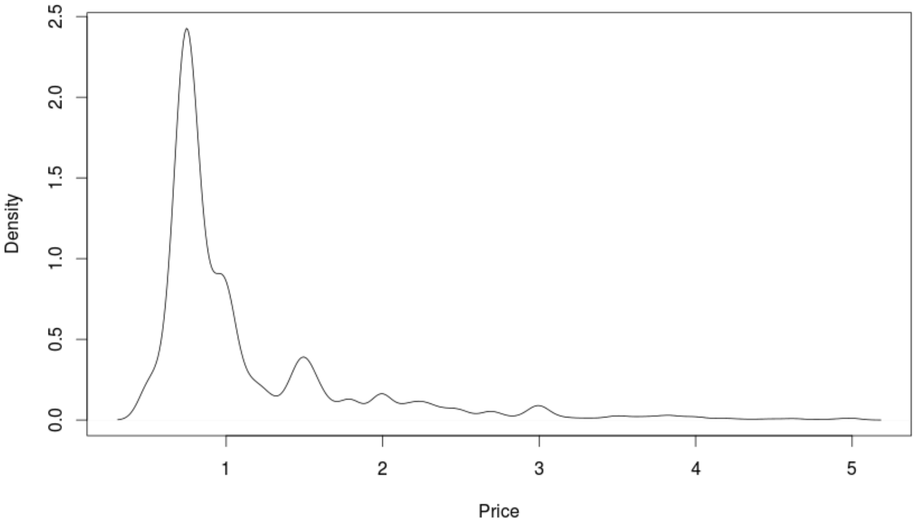


Figure 3: Probability Density Function (PDF) of non-free Arcade apps with a price ranging between 0.5 and 5 euros. They are 99% of non-free Arcade apps.

higher is the hidden "price" paid by users in terms of privacy leakage to compensate the low price in the store.

The purpose of this paper is to provide a first quantification to that hidden price on the basis of the observation made on the prices of non-free applications available in the store. The proposed method is based on the process depicted in figure 5 and is made of the following steps:

S1 We assume that given a specific set of permissions a certain amount of private information is disclosed.

S2 We use our dataset, to generate a model capable to infer the price of the applications that request the access to a specific set of permission.

S3 Under the assumption made in S1, this model allows us to establish a relationship between the amount of private information disclosed and a price.

In other words, we use the inferred price as a proxy to attribute an economic value to the private information disclosed by users when they grant the access to such set of permissions. Notice that according to H1, the price attributed to the private information should compensate the price in the store that is predicted by the model However, irrespectively from

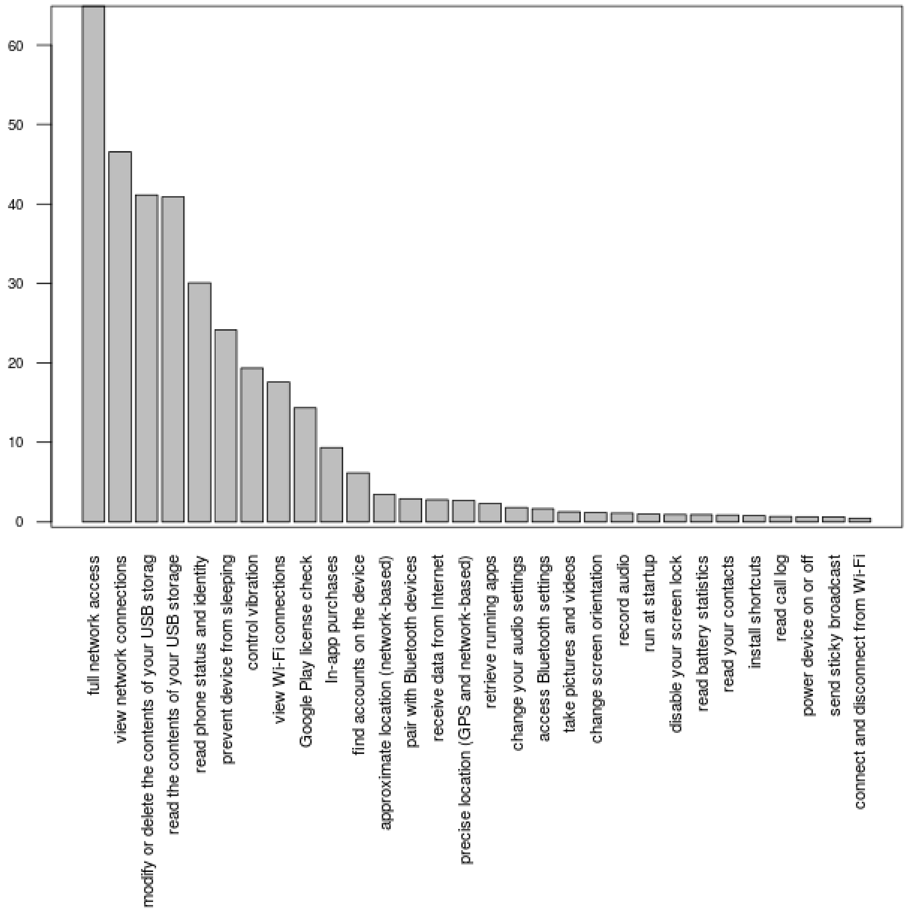


Figure 4: Frequency of permission of non-free Arcade apps with a price ranging between 0.5 and 5 euros. For the sake of readability, only the first 30 permissions in the rank are shown.

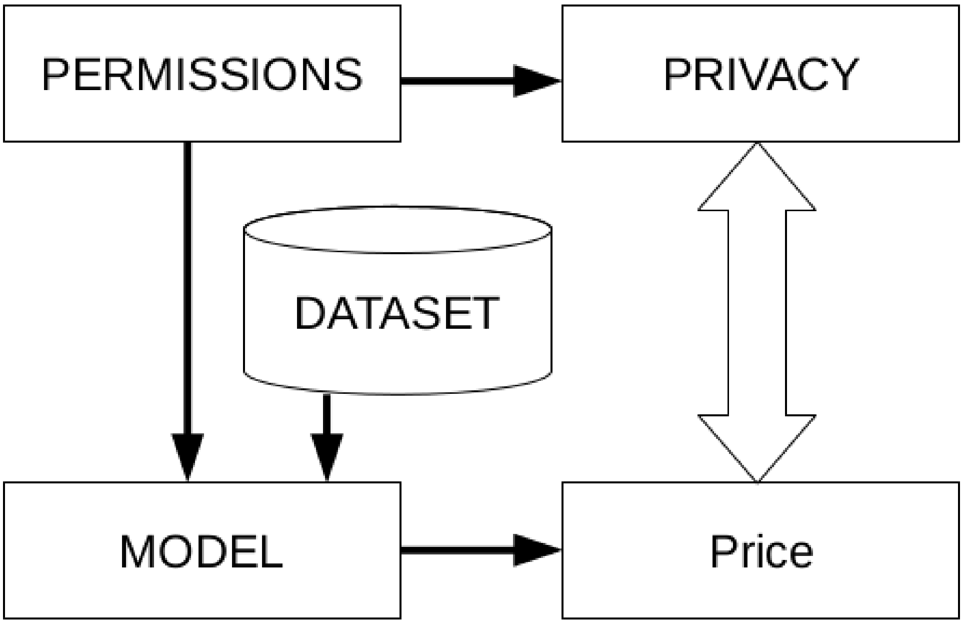


Figure 5: A picture of the process to estimate the price of private information.

the considered relationship between the price in the store and the price of private information, a key element to validate our hypothesis is a good model capable to accurately predict the price in the store.

We started observing the heatmap shown in figure 6 that correlates the number of permissions used by an application with its price. Notice that free applications require a bigger number of permissions. This seems to support the evidence that free applications are indirectly "paid" by users providing access to an higher number of permissions, thus potentially disclosing more private information.

However, such observation, does not allow us to quantify the price associated with a given set of permission as shown in figure 5. To such purpose, we use a linear regression model [6] capable to estimate the price having in input the set of requested permissions as independent dummy variables.

4.2 Linear Regression Model

The Linear Regression Model is described by the following equation

$$Price = \sum w_i P_i + Intercept \tag{1}$$

In which $P_i = 1$ if the permission P_i is requested by the application and zero otherwise. The weights w_i and the *Intercept* have been evaluated in R [2] using the *lm* function.

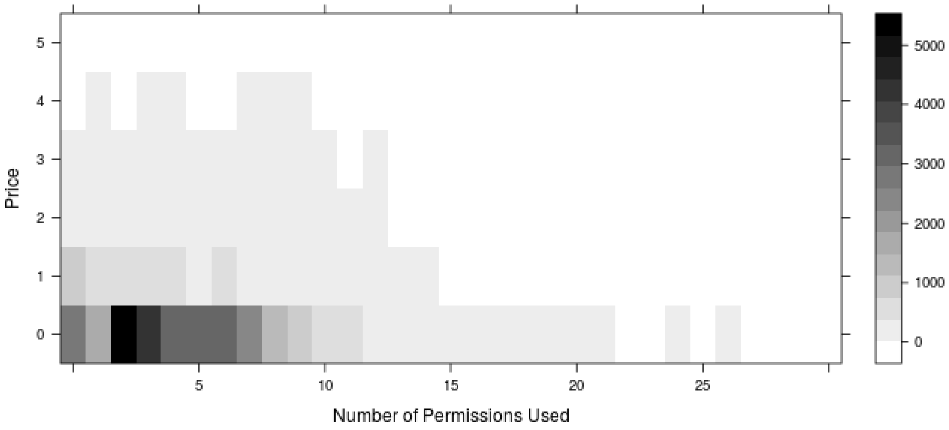


Figure 6: Heatmap correlating the number of permission used and the price of Arcade applications. Notice that free applications require a bigger number of permissions.

Among the 266 permissions given in input to the model, only 84 have been considered statistically significant by the model. Among these, the ones with the lower p-values (i.e. Probability the variable is NOT relevant) are:

Table 2: The most relevant independent variables (i.e. permissions in the model.

Permission	p-values
Read phone status and identity	$5.90e-05$
Prevent device from sleeping	$1.37e-05$
Modify/delete internal media storage contents	$6.42e-07$
Google Play license check	$< 2e-16$
In-app purchases	$2.32e-05$

While it is not clear why *Prevent device from sleeping* would be an issue in terms of disclosure of private information, the other permissions are clearly related to user's privacy.

A crucial point is to what extent the model can accurately predict the price. A first estimation of the model accuracy can be made observing the residuals. The residuals of the linear regression model are the difference between the observed data of the dependent variable *Price* and the fitted values $\widehat{Price}$, namely

$$Residual = Price - \widehat{Price} \tag{2}$$

Residuals can be seen as the error made by the model prediction in

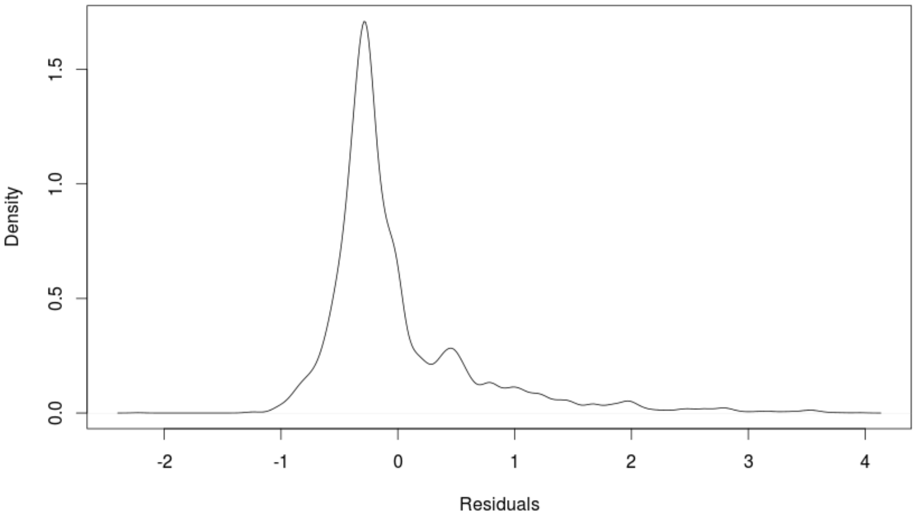


Figure 7: PDF of residuals.

estimating observed data and thus provide a good indication on the accuracy of the model. The density of residuals is shown in figure 7.

This figure shows that the residuals are concentrated between -1 and 1 euro with a standard error of 0.6778; Considering that the price for the applications ranges between 0.5 and 5 euros the error can be significantly high.

This is further confirmed by the figure 8 in which the Price of each application is plotted vs the corresponding residual. This figure shows that the residuals are between 0 and about 1 euro in defect for the vast majority of the cases. While this is a relatively small error for prices bigger than 2-3 euros, it becomes significant for smaller prices that unfortunately are the majority (see figure 3).

We conclude this analysis plotting the Empirical Cumulative Distribution Function (ECDF) of the absolute value (errors in excess and in default) of the residuals normalised by the price, namely:

$$|\frac{Price - \widehat{Price}}{Price}| \tag{3}$$

This metric is an indication of the relative error with respect to the real price of the application. As can be observed in figure 9, in 60% of the cases the model estimates the price with an error of at most 50%. However, for the other 40% of the cases, the error can be from 50% to 250%.

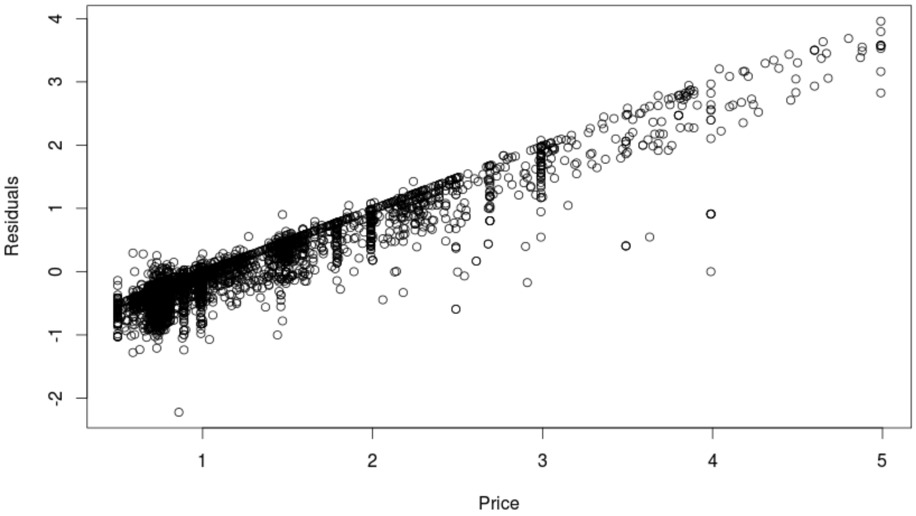


Figure 8: Price vs Residuals.

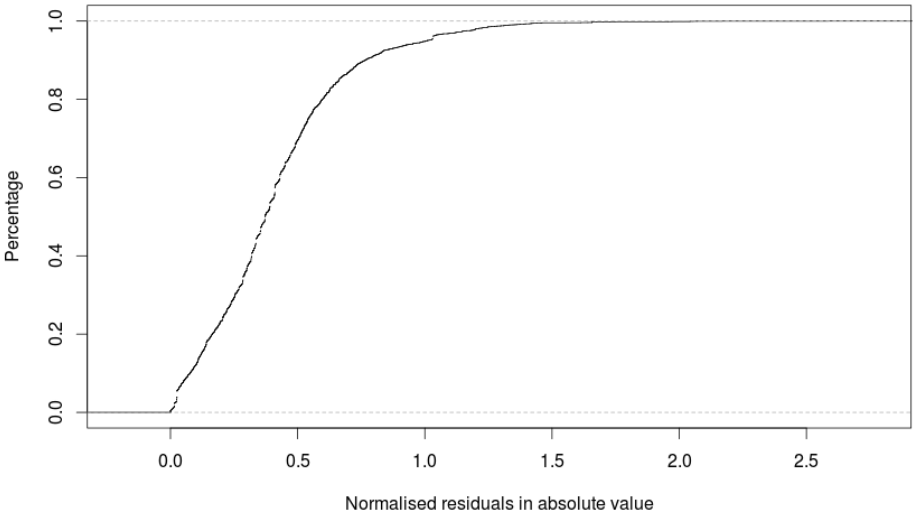


Figure 9: ECDF of the normalised residuals in absolute value.

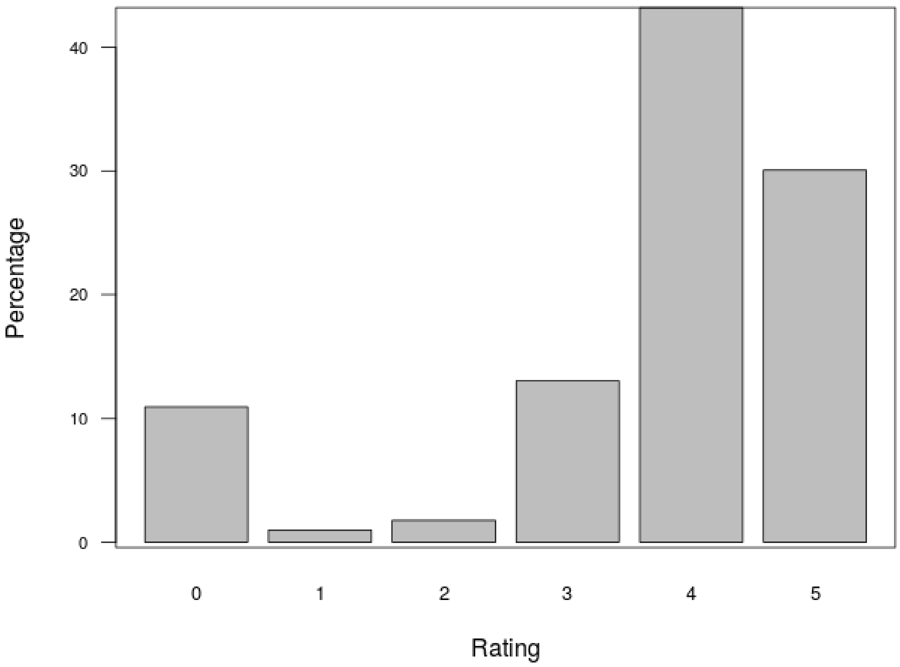


Figure 10: Frequency of ratings of Arcade apps in percentage. Most applications get a good rating (> 3).

5 Estimate the price using the Ranking

In this section, we explore to what extent the ranking of an application can be used to predict its price. Even if the ranking is not strictly related with private information as in the case of the grated permissions, here we are interested in understanding whether it can allow use to better estimate the price of an application.

Figure 10 shows the distribution of the rating; notice that most applications get a good rating (> 3).

5.1 Linear Regression Model

In this case the Linear Regression Model evaluated again in R using the *lm* function is described by the following equation

$$Price = 0.038 \cdot Rating + 1.03 \tag{4}$$

Each point in the figure 11 is the rating and the corresponding price of an application in the dataset. The red line, is the model represented by equation 4. As can be observed from this figure, the linear model does not

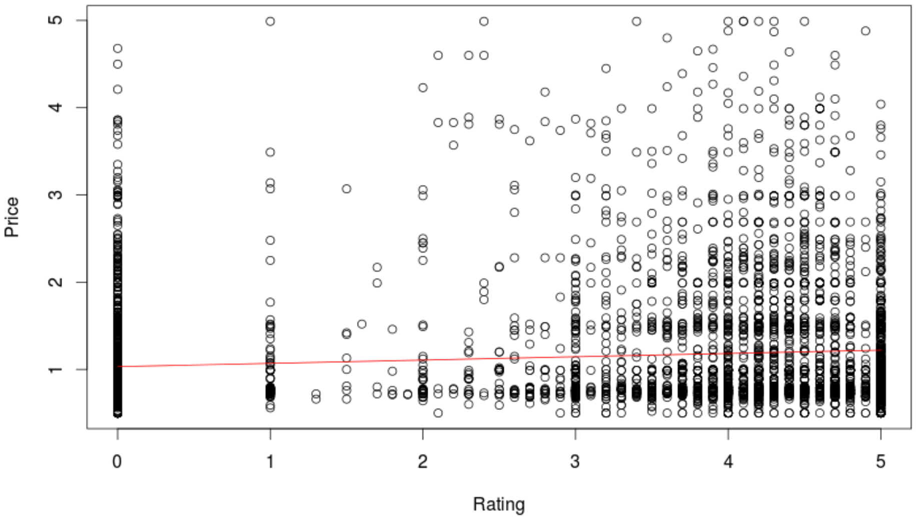


Figure 11: Each point in the figure is the rating and the corresponding price of an application in the dataset. The red line, is the linear model in equation 4.

properly capture the structure of the dataset and thus its prediction power is limited.

Even if the plot of Price vs Residuals in figure 12 shows that in this case the variance is lower, the ECDF of the absolute value of the residuals normalised shows that again in 60% of the cases the model estimates the price with an error of at most 50% (see figure 13). However, for the other 40% of the cases, the error is slightly better than in the previous case and is bound to 150%.

6 Discussion and Future Work

In this paper we estimated the price using a multivariate linear regression model providing in input as independent variables all the permissions in the dataset. An interesting question is to explore whether non linear regression models, such as model trees [6], can improve the accuracy of the prediction. Furthermore, a careful selection of the permissions can likely provide better or more significant results. In terms of privacy, the permission that allows an application to access the light is not as critical as the permissions that grants access to the address book or to the network.

According to our hypothesis, free applications require more permissions that allows developer to access more private information and possibly make

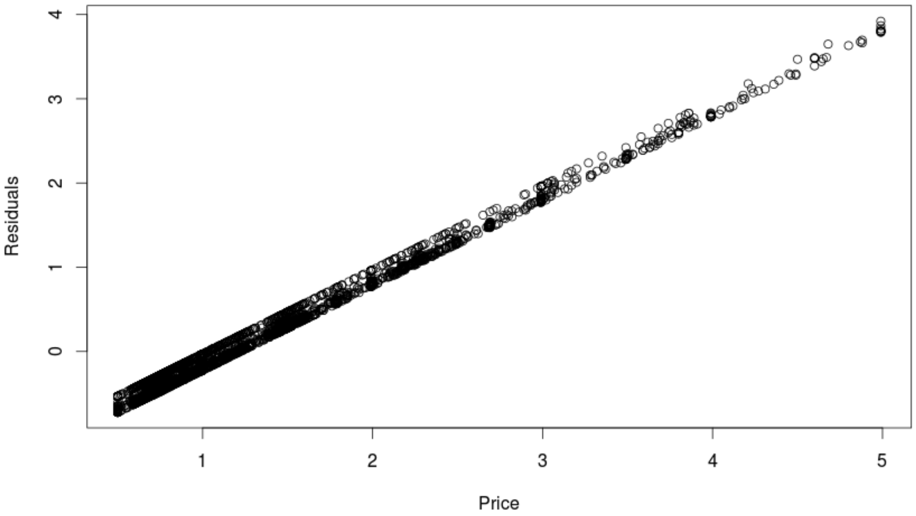


Figure 12: Price vs Residuals.

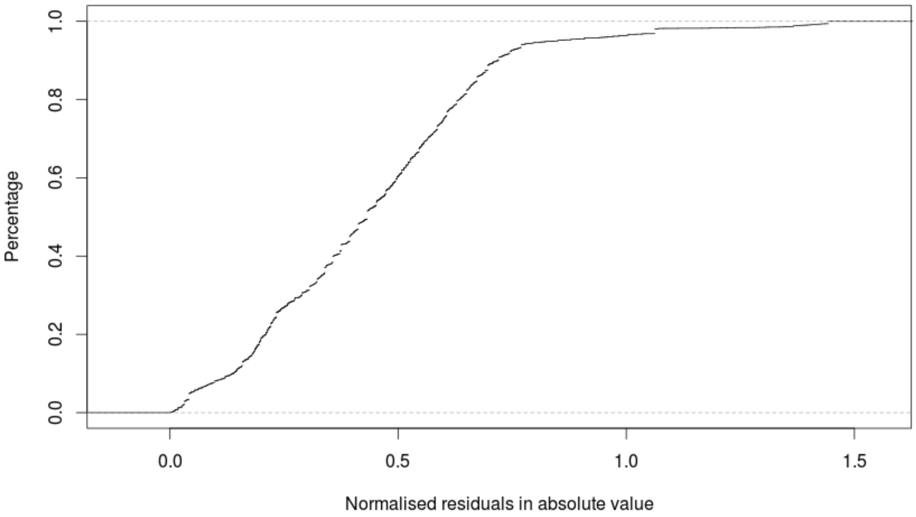


Figure 13: ECDF of the normalised residuals in absolute value.

money out of it. However, another possible explanation is that developers simply take less care in the selection of free applications' required permissions in order to speed up the development of a product that is going to be released free-of-charge. On the contrary, they more carefully select requested permissions when an application is non-free to eventually give users a higher quality app. This is also supported by the observation made by the authors of [5] "We investigate the causes of overprivilege and find evidence that developers are trying to follow least privilege but sometimes fail due to insufficient API documentation"

In the future, we plan to extend the analysis to all the 1,135,700 applications in our dataset. As a roadmap for the future work, we foresee the following main steps:

- Automatically classify free vs non-free applications, possibly considering the application in clusters defined by the category
- Accurately predict the price of non-free application
- Study the relationship between predicted price and private information value (see H1 in section 4)
- Design and implement a tool to provide quantitative indication to the users on the economic value of their private information

References

[1] Google play help web site - About app permissions. `https://support.google.com/googleplay/answer/6014972?hl=en`.

[2] The R project for statistical computing web site. `http://www.r-project.org/`.

[3] Selenium HQ browser automation web site. `http://www.seleniumhq.org/`.

[4] Alastair R. Beresford, Andrew Rice, Nicholas Skehin, and Ripduman Sohan. Mockdroid: Trading privacy for application functionality on smartphones. In *Proceedings of the 12th Workshop on Mobile Computing Systems and Applications*, HotMobile '11, pages 49–54, New York, NY, USA, 2011. ACM.

[5] Adrienne Porter Felt, Erika Chin, Steve Hanna, Dawn Song, and David Wagner. Android permissions demystified. In *Proceedings of the 18th ACM Conference on Computer and Communications Security*, CCS '11, pages 627–638, New York, NY, USA, 2011. ACM.

[6] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

[7] Jon Kleinberg, Christos H. Papadimitriou, and Prabhakar Raghavan. On the value of private information. In *Proceedings of the 8th Conference on Theoretical Aspects of Rationality and Knowledge*, TARK '01, pages 249–257, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[8] Ilias Leontiadis, Christos Efstratiou, Marco Picone, and Cecilia Mascolo. Don't kill my ads!: Balancing privacy in an ad-supported mobile application market. In *Proceedings of the Twelfth Workshop on Mobile Computing Systems &#38; Applications*, HotMobile '12, pages 2:1–2:6, New York, NY, USA, 2012. ACM.

[9] R. Stevens, C. Gibler, J. Crussell, J. Erickson, and H. Chen. Investigating user privacy in android ad libraries. In IEEE Symposium on Security and Privacy 2012 Workshops, 2012.

[10] Thein Than Tun, A.K. Bandara, B.A. Price, Yijun Yu, C. Haley, I. Omoronyia, and B. Nuseibeh. Privacy arguments: Analysing selective disclosure requirements for mobile applications. In *Requirements Engineering Conference (RE), 2012 20th IEEE International*, pages 131–140, Sept 2012.