Which Apps have Privacy Policies?
An analysis of over one million Google Play Store apps***

Peter Story[0000−0002−3174−2563], Sebastian Zimmeck[0000−0002−2500−2681], and Norman Sadeh[0000−0003−4829−5533]

Carnegie Mellon University, Pittsburgh PA 15213, USA
{pstory,szimmeck,ns1i}@andrew.cmu.edu

Abstract Smartphone app privacy policies are intended to describe smartphone apps' data collection and use practices. However, not all apps have privacy policies. Without prominent privacy policies, it becomes more difficult for users, regulators, and privacy organizations to evaluate apps' privacy practices. We answer the question: "Which apps have privacy policies?" by analyzing the metadata of over one million apps from the Google Play Store. Only about half of the apps we examined link to a policy from their Play Store pages. First, we conducted an exploratory data analysis of the relationship between app metadata features and whether apps link to privacy policies. Next, we trained a logistic regression model to predict the probability that individual apps will have policy links. Finally, by comparing three crawls of the Play Store, we observe an overall-increase in the percent of apps with links between September 2017 and May 2018 (from 41.7% to 51.8%).

Keywords: privacy · privacy policy · smartphone · smartphone apps

1 Introduction

The Google Play Store makes over a million apps accessible to Android users in the US [29]. Many apps collect location details, contact information, phone numbers, and a variety of other data from their users [20]. Oftentimes, the collected data is not only leveraged for the apps' main functionalities but also for other purposes, most notably, to serve advertisements and for analytics. The notice and choice paradigm prescribes that app developers should notify their users of how they collect, use, share, and otherwise process user information in

* This study was supported in part by the NSF Frontier grant on Usable Privacy Policies (CNS-1330596 and CNS-1330141) and a DARPA Brandeis grant on Personalized Privacy Assistants (FA8750-15-2-0277). The US Government is authorized to reproduce and distribute reprints for Governmental purposes not withstanding any copyright notation. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the NSF, DARPA, or the US Government.

** To appear at the Annual Privacy Forum 2018.

their privacy policies. The promises contained in these policies are enforceable by privacy regulators and are of interest to privacy-focused organizations and researchers.

In the absence of comprehensive federal legislation in the US, the California Online Privacy Protection Act requires online services that collect personally identifiable information to post a policy.[1] A similar requirement is contained in Delaware's Online Privacy and Protection Act.[2] Further, the Federal Trade Commission's Fair Information Practice Principles call for consumers to be given notice of an entity's information practices before any personally identifiable information is collected [16]. The Children's Online Privacy Protection Act makes policies mandatory for apps directed to or known to be used by children.[3]

The Google Play Store gives app developers the option to include links to their privacy policies on their Play Store pages. However, in three separate crawls of apps we found that only 41.7% (August 28 through September 2, 2017—in the following "First Crawl"), 45.2% (November 29 through December 2, 2017—in the following "Second Crawl"), and 51.8% (May 11 through May 15, 2018—in the following "Third Crawl") have such links. While there appears to be an upward trend, these percents are relatively low, especially, as they include links for apps that are legally required to disclose their practices in privacy policies (most notably, apps that are subject to the Children's Online Privacy Protection Act [15]).

In this study we aim to identify app features that are associated with whether an app links to a privacy policy or not. To that end, we offer the following contributions:

1. We present an in-depth exploratory analysis of features associated with whether apps have privacy policies (§ 4). Among other findings, our analysis reveals that only 63.1% of apps which describe themselves as sharing their users' locations link to privacy policies.
2. We design a logistic regression model which quantifies the associations between policy links and other app features (§ 5). For example, our model indicates that an app with a developer address in Germany has greater odds of having a policy link than an app without country information.
3. We discuss how our work might be useful to government regulators, privacy organizations, and researchers (§ 6). In particular, we provide suggestions about how our techniques can be used to prioritize regulatory enforcement actions, evaluate the relative merits of individual app features, and observe trends over time.

2 Related Work

We are aware of several previous studies examining privacy policy occurrence in the app ecosystem. Our work goes beyond these studies by analyzing orders

[1] Cal. Bus. & Prof. Code §22575(a).
[2] Del. Code Tit. 6 §1205C(a).
[3] 16 CFR §312.4(d).

of magnitude more apps and by employing more scalable analysis techniques. Sunyaev et al. analyzed the presence of privacy policies for the most popular health-related Android and iOS apps [30]. In addition to following links from the Play Store, they searched for policies on developers' websites and Google. They found that only 22.7% of the Android apps with the most ratings in the Health and Fitness and Medical categories had privacy policies. Blenner et al. analyzed the privacy polices and practices of a random sample of diabetes-related Android apps [3]. They found that only 19% of apps had privacy policies. Different from these previous studies, we conclude that a substantially higher percent of apps in the Health and Fitness and Medical categories linked to privacy policies from their Play Store pages (45.6% and 45.0%, respectively). Our finding suggests that it is now more common for app developers to link to privacy policies (§ 6) than at the time of the previous studies.

Instead of gathering data from the Play Store directly, Balebako et al. interviewed and surveyed US app developers about how they protect the privacy of their users [2]. 57.9% of the developers they surveyed reported hosting a privacy policy on their website. In comparison, we found that 64.0% of apps with US mailing addresses link to privacy policies, which suggests an increase over time. Balebako et al. found a generally positive relationship between company size and whether companies have privacy policies.

In the closest work to ours, Zimmeck et al. analyzed 17,991 free Android apps for features that identify apps with privacy policy links [33]. They found a number of features useful for predicting whether an app has a privacy policy: recent app update, small or large number of installs, Editors' Choice or Top Developer badges, in-app purchase offers, and Entertainment Software Rating Board (ESRB) content ratings [12] appropriate for younger audiences. However, they also found that apps in the Comics, Libraries & Demo, Media & Video, and Personalization categories had particularly low percents of policies. In this report, we not only repeat the analysis of these features,[4] but we go beyond their examination in multiple dimensions. First, we collected the metadata of a much larger set of apps. We also take into account features that were not analyzed by Zimmeck et al., including apps' ESRB content descriptors, prices, and developers' home countries. In addition, we train a logistic regression model which considers all these features together. Further, the repetition of our analysis gave us insight into how the app population changes over time (§ 6).

A number of other researchers have performed analyses at the scale of the entire Google Play Store, however, not for purposes of predicting whether apps have privacy policy links. In particular, d'Heureuse et al. used multiple crawling techniques to explore the app ecosystem, including browsing by category, by related apps, and by searching [11]. One notable finding was that only 46% of apps in the Google market were discoverable solely by following links to related apps. However, whether this finding is still true today is unclear. In our First and Second Crawls, we found over a million apps by following links to related

[4] We were unable to consider the Top Developer badge in our analysis because it is no longer displayed on the Play Store [13,26].

apps. By the time of our Third Crawl, only about 179K apps could be found when just this technique was used (§ 3). Viennot et al. also used searching techniques to discover apps on the Google Play Store [31]. Wang et al. analyzed the privacy characteristics of Play Store apps [32]. However, they did not consider whether apps linked to privacy policies. Different from prior work, we focus on the prevalence of links to privacy policies.

This report is also informed by our earlier work in the field of smartphone app privacy, performed as part of the Usable Privacy Policy Project [28].[5] In particular, Lin et al. used crowdsourcing to detect unexpected uses of data by smartphone apps [25]. Kelley et al. demonstrated that alternate presentations of apps' privacy-related behavior can impact which apps users install [23]. Lin et al. clustered users based on their app privacy preferences [24]. Almuhimedi et al. used nudges to encourage users to customize their smartphone permission settings [1].

3 Methodology

It is our goal in this study to find features that predict the occurrence of privacy policy links for apps. The features we examined were obtained from apps' Play Store metadata and include, among others, the average rating assigned by reviewers, how many times the app was installed, and the Play Store categories the app belongs to.

Starting with a randomly selected app (com.foxandsheep.littlefox), we recursively followed links to related apps. This technique is relatively resource efficient: on a single virtual server,[6] our crawls all completed in less than a week. Our First and Second Crawls used only this recursive technique. However, by the time of our Third Crawl, only about 179K could be found when just this technique was used. We think this is because Google altered the algorithms which recommend related apps. Consequently, for the Third Crawl we seeded the database with the app identifiers collected by the Second Crawl. Using these techniques, we retrieved the metadata associated with $n = 1,423,450$ apps (First Crawl), $n = 1,163,622$ apps (Second Crawl), and $n = 1,044,752$ (Third Crawl). Unless otherwise noted, all statistics and results described herein refer to the Second Crawl. Also, note that our results refer to the US Play Store.

For each feature, we perform two types of analyses (§ 4). First, we evaluate the relative occurrence of the different values of a feature (e.g., for the install count we leverage the install ranges given on the Play Store and evaluate the percent of apps that were installed 1–5 times, 5–10 times, etc.). Second, we examine the relative occurrence of apps with privacy policy links at different feature values (e.g., for apps that were installed 1–5 times, 49.5% have a privacy policy; for apps that were installed 5–10 times, 47.7% have a privacy policy; etc.).

[5] https://www.usableprivacy.org, accessed: May 20, 2018.

[6] Our virtual server had four Intel Xeon E5-2640 CPU cores at 2.50GHz and 8GB of RAM.

Next, based on the results of our feature analysis, we build and evaluate a logistic regression model for predicting whether apps link to privacy policies from their Play Store pages (§ 5).

It is a limitation of our approach that some apps may not have a link to their policy on their Play Store page, but rather provide such a link in another place (e.g., inside their code). However, using privacy policy links on the Play Store as proxies for actual policies is not unreasonable since regulators requested that app publishers post such links [22,14], and app store owners obligated themselves to provide the necessary functionality [21]. Furthermore, Zimmeck et al. found that of apps which didn't link to their policy from the Play Store, only 17% of apps provided their policies somewhere else [33]. Also, in order for the notice and choice model to be effective, users should be able to examine an app's privacy policy before they install it. In future work, we will go beyond this assumption by seeking links to privacy policies within the apps themselves using static code analysis.

Another limitation is that our recursive crawling technique may not discover all the apps on the Play Store. However, based on the large number of apps included in our crawls, we estimate that our crawls covered the vast majority of apps that a typical user would encounter.

4 Exploratory Data Analysis of Potentially Relevant Features

We find that 45.2% of apps link to their privacy policy from their Play Store page. We now seek to explore the features that predict such occurrence. In the following we examine two types of features: native Play Store features (§ 4.1), such as an app's install range on the Play Store, and ESRB features (§ 4.2), such as ESRB content ratings.

4.1 Play Store Features

Country (Figure 1) While Google does not require that developers display their countries of origin, some post a contact mailing address. With a few steps of pre-processing we were able to determine the countries of 17.2% of apps. First, we extracted the country from each address using the `pypostal` library.[7] Note that we skip addresses which do not explicitly include a country.[8] We cleaned the data by consolidating abbreviations and alternate spellings of countries using the `pycountry` library.[9] Further, we wrote custom mappings for all other countries except for those with fewer than 30

[7] https://github.com/openvenues/pypostal, accessed: May 20, 2018.

[8] Consequently, the relative frequencies shown in Figure 1 should be interpreted cautiously. For example, it would not be safe to assume that there are more developers from India than from the US as developers from India may possibly include the country in their address more frequently than developers from the US.

[9] https://bitbucket.org/flyingcircus/pycountry, accessed: May 20, 2018.

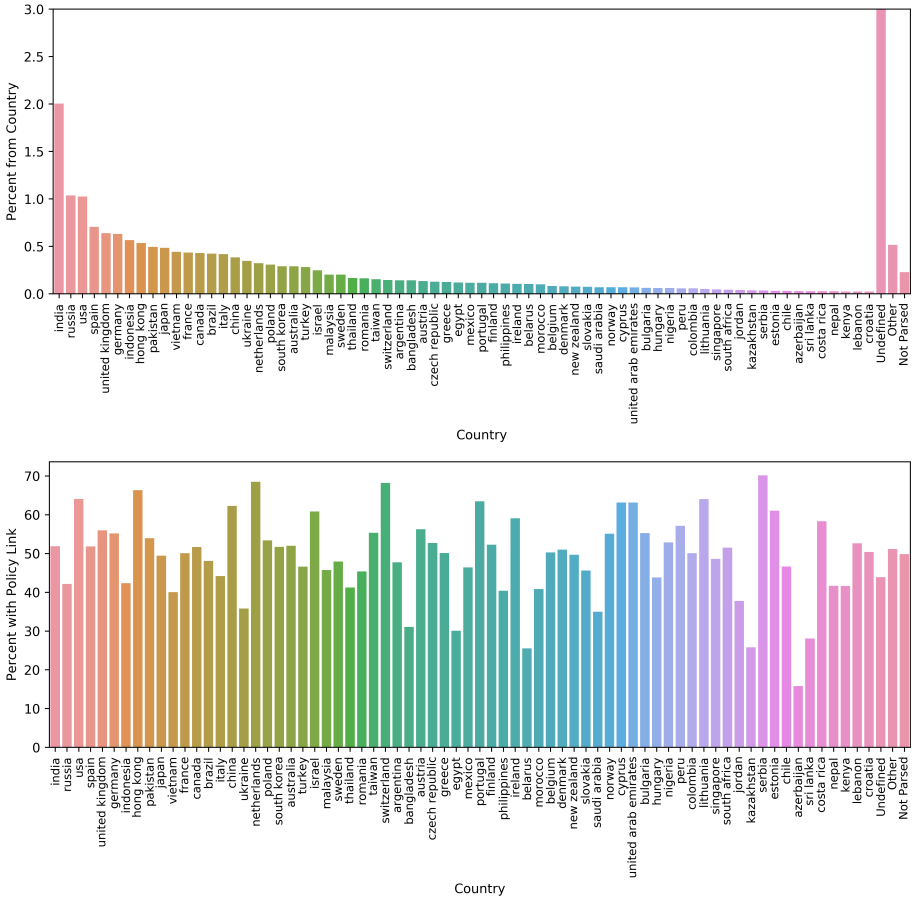

Figure 1. Percent of apps per country (top) and respective privacy policy percent (bottom).

associated apps. These 0.2% of apps are shown as "Not Parsed" in Figure 1. In total we were able to fully extract the countries for 17.2% of apps. The remaining 82.5% of apps either did not have an address on the Play Store or our technique was unable to extract a country from the address that was posted (shown as "Undefined"). Finally, countries associated with fewer than 250 apps were combined in the "Other" category (unless they were already included in the "Not Parsed" category).

As Figure 1 shows, there are many developers publishing apps on the US Play Store from countries other than the US (most notably, from India and Russia). As discussed later (§ 5.2), we found that some country features affect the odds of apps having privacy policies.

It should be noted that we were only able to determine the country for 17.2% of apps; this result is based on Google's decision to not require developers to post a mailing address. Also, the addresses which are posted do not have a consistent format, and many addresses are given without country information. However, the country data we did extract were still salient, since many countries were retained in our logistic regression model.

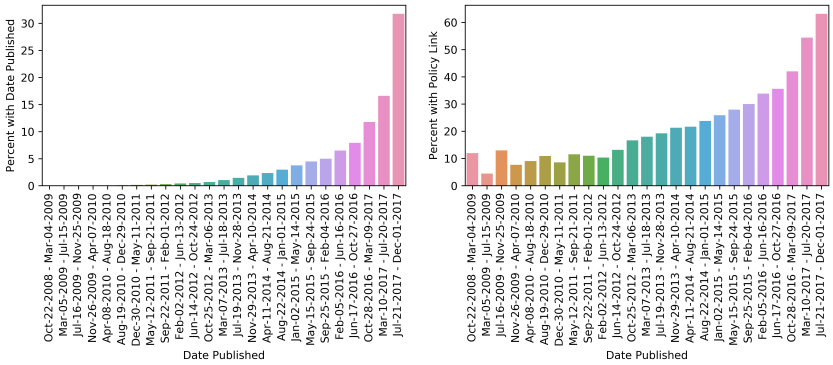

Figure 2. Percent of apps by date published (left) and respective policy percent (right).

Date Published (Figure 2) Apps' Play Store pages display the date when its latest version was published. If the app was never updated, it will be the date when it was first released. Figure 2 shows a distribution which is skewed to the left indicating that most apps have been published recently. Similar to earlier results [33] our analysis appears to show that apps published more recently are more likely to have privacy policies.

Editors' Choice Google assigns Editors' Choice badges to "apps and games with the best experiences on Android" [4]. Just 621 apps have the Editors' Choice badge, of which 93.1% have a privacy policy. As only 45.2% of apps without such badge have privacy policies, it appears to be a strong signal. However, given the small number of apps that have a badge, its impact overall is rather limited.

Install Ranges (Figure 3) The Play Store does not display the exact number of installs of apps. Instead, at the time of our First and Second Crawls it displayed ranges of installs (e.g., 1–5 installs, 5–10 installs, etc.).[10] Figure 3 shows that the distribution of app install ranges has a long tail. In particular, it should be noted that there are only a few apps with billions of installs. Many of those apps have privacy policies. However, even apps with very few installs often have privacy policies. In fact, beginning with apps having

[10] By the time of our Third Crawl, the Play Store had changed the display of the install ranges and started showing only their smallest values, e.g., 1+ installs, 5+ installs, etc.

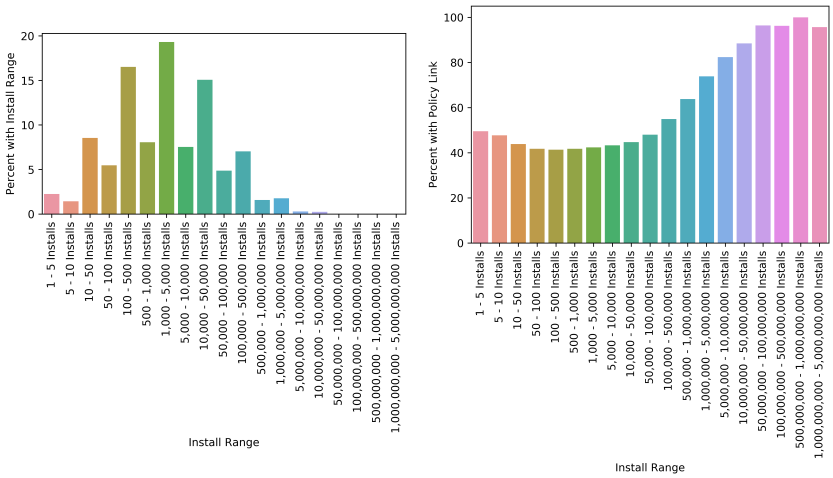

Figure 3. Percent of apps per install range (left), and respective policy percent (right).

1–5 installs the percent of apps with policies decreases to a low point at 100–500 installs, then generally increases from there. This finding confirms a trend that was observed earlier: apps with relatively high and low install ranges are more likely to have privacy policies than apps with medium install ranges [33]. A hypothesis provided by Zimmeck et al. was that apps with fewer installs were more recently published and hence more likely to aim for privacy compliance [33].

Play Store Category (Figure 4) Apps on the Play Store are organized by category. A given app can be part of multiple categories. Figure 4 shows how the percent of apps with policies differs by category. In particular, it can be observed that some of the most popular categories—BOOKS_AND_REFERENCE, EDUCATION, and ENTERTAINMENT—are among those with the lowest prevalence of policies. Further, notice that 100% of apps in the FAMILY_ categories have policies. The reason for this complete coverage is Google's management of those categories in the Designed for Families program [17] that requires all apps to have a privacy policy.

Price (Figure 5) The price of an app is the cost associated with installing that app, without considering in-app purchases. Figure 5 shows that 99.5% of apps are either free or can be purchased for $5 or less. Although there does not seem to be an obvious relationship between an app's price and whether it has a policy, price turned out to be a significant feature in our model (§ 5.2).

Rating Count (Figure 6) Play Store users can rate apps on a scale of one to five (worst to best). The rating count is the number of ratings an app has received. Figure 6 shows that the distribution of rating counts is strongly skewed: most apps have only a few ratings but some have much higher counts. 12.5% of apps have no ratings. Fewer than 9% of apps have more than 1,000

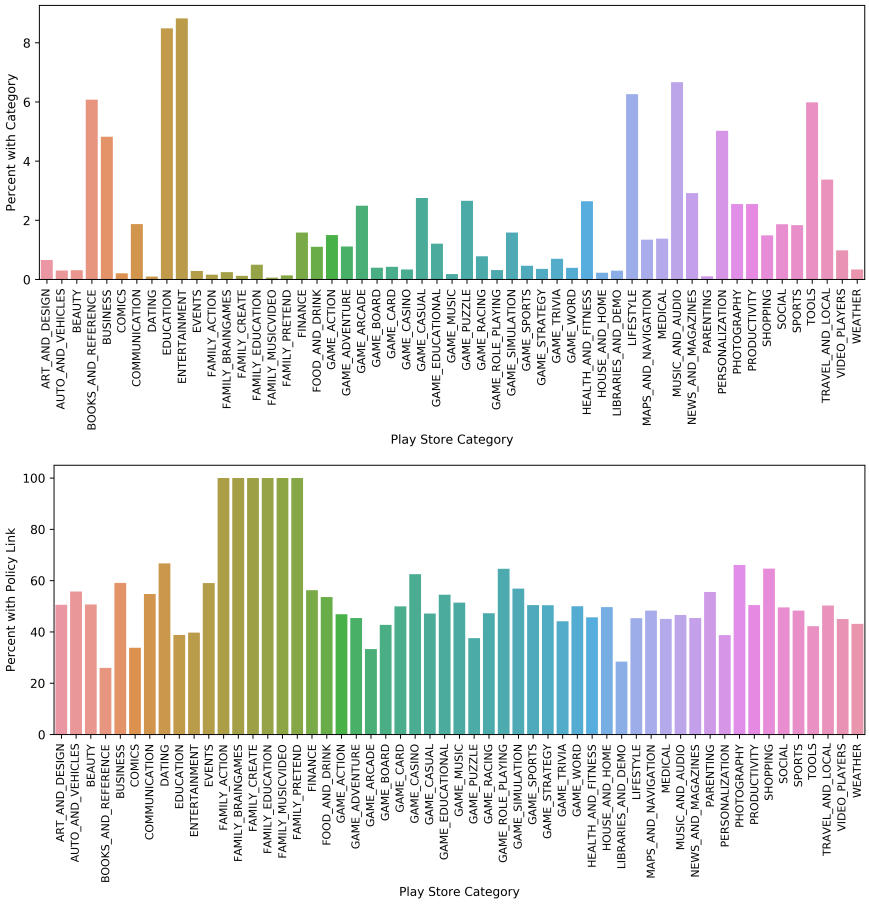

Figure 4. Percent of apps per Play Store category (top) and respective policy percent (bottom).

ratings. Figure 6 appears to show that apps become more likely to have privacy policies as their number of ratings increases. Our logistic regression analysis confirmed this observation (§ 5.2). The trend seems similar to the trend for install ranges (see Figure 3).

Rating Value (Figure 7) The rating value is the average of all its user ratings. Figure 7 shows the percent of apps with different average rating values. The peaks at 1, 2, 3, 4, and 5 might be caused by apps that have only a few ratings: in those cases, it is more likely that the average rating will be a whole number. While Figure 7 does not show an obvious connection between rating value and whether apps have policies, our logistic regression analysis actually discovered a nonlinear relationship between the two (§ 5.2).

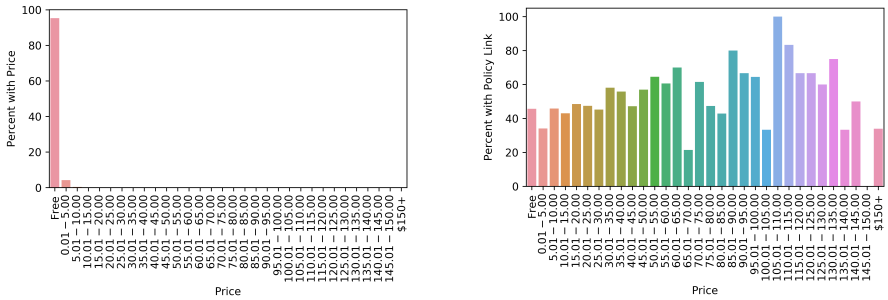

Figure 5. Percent of apps per price category (left) and respective policy percent (right).

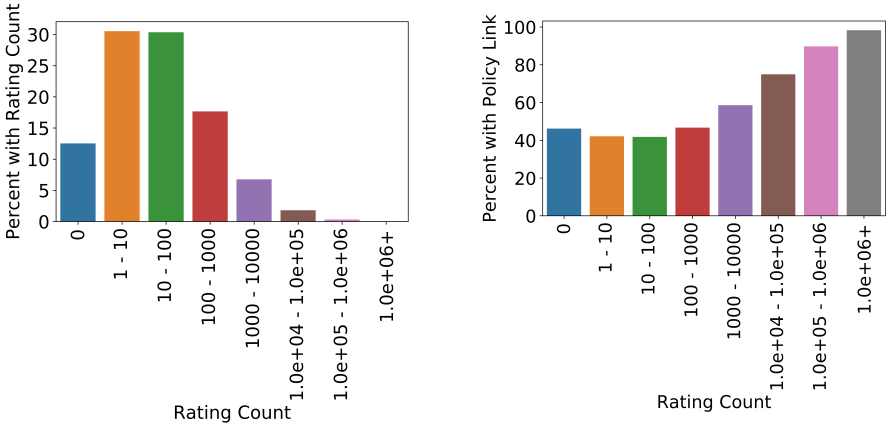

Figure 6. Percent of apps per rating category (left) and their respective policy percent (right). Note the use of log-scales on the x-axes.

4.2 ESRB Features

Google provides a questionnaire that developers can use to describe the content of their app [18,19]. The answers to this questionnaire are used to generate an app's ESRB content rating, its ESRB content descriptors, and its interactive elements. All this information is displayed to users on the US Play Store [12].

ESRB Content Rating (Figure 8) ESRB content ratings define the age categories an app is appropriate for, and every app has exactly one such rating. Figure 8 shows that over 84% of apps in our sample are rated as suitable for EVERYONE. There are comparatively few apps with other ratings. In particular, we found only 44 apps with the ADULTS rating. It can be observed that the UNRATED apps appear to be much less likely to have policies. It is encouraging to see that TEEN-rated apps have the highest policy percent. However, it is also true that many apps rated EVERYONE 10+ do not have policies.

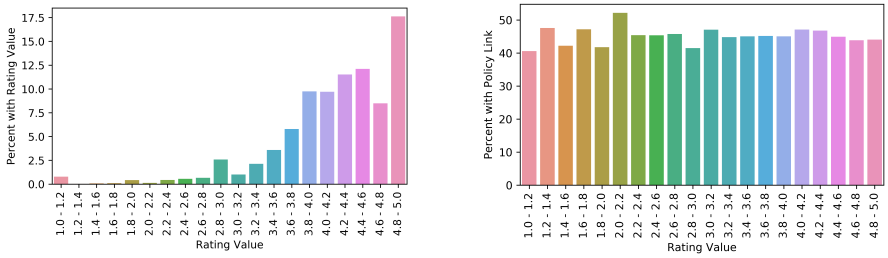

Figure 7. Percent of apps per average rating category (left) and respective policy percent (right).

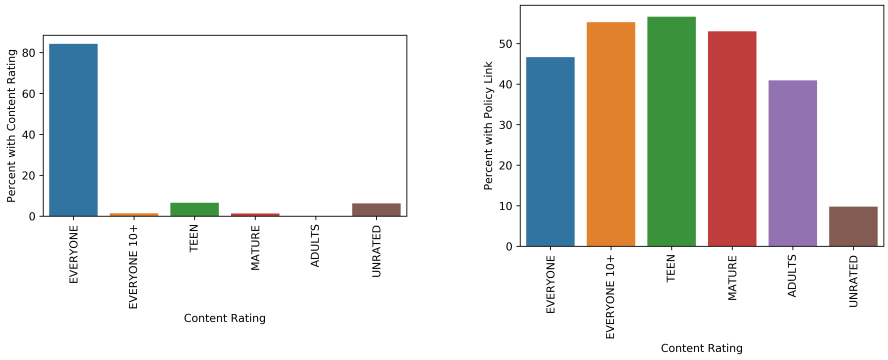

Figure 8. Percent of apps per ESRB content rating (left) and respective policy percent (right).

ESRB Content Descriptors (Figure 9) Content descriptors describe app content that is potentially objectionable to certain users. Figure 9 shows the relative frequencies of different content descriptors and how the percent of apps with policies differs by descriptor. Only 12% of apps have one or more content descriptors. The Warning descriptor is by far the most used one.[11] It also is the content descriptor with the second-lowest policy percent. The Mild Sexual Themes descriptor is only used by one app. Thus, its 0% policy coverage is of very limited relevance.

Interactive Elements (Figure 10) ESRB interactive elements describe five other characteristics of apps that are of potential interest to users. 18% of apps have one or more interactive element. Our logistic regression analysis found that all interactive elements except Unrestricted Internet were associated with an increase in the odds that an app would have a privacy policy (§ 5.2).

[11] Note that the full Warning descriptor reads "Warning - content has not yet been rated. Unrated apps may potentially contain content appropriate for mature audiences only."

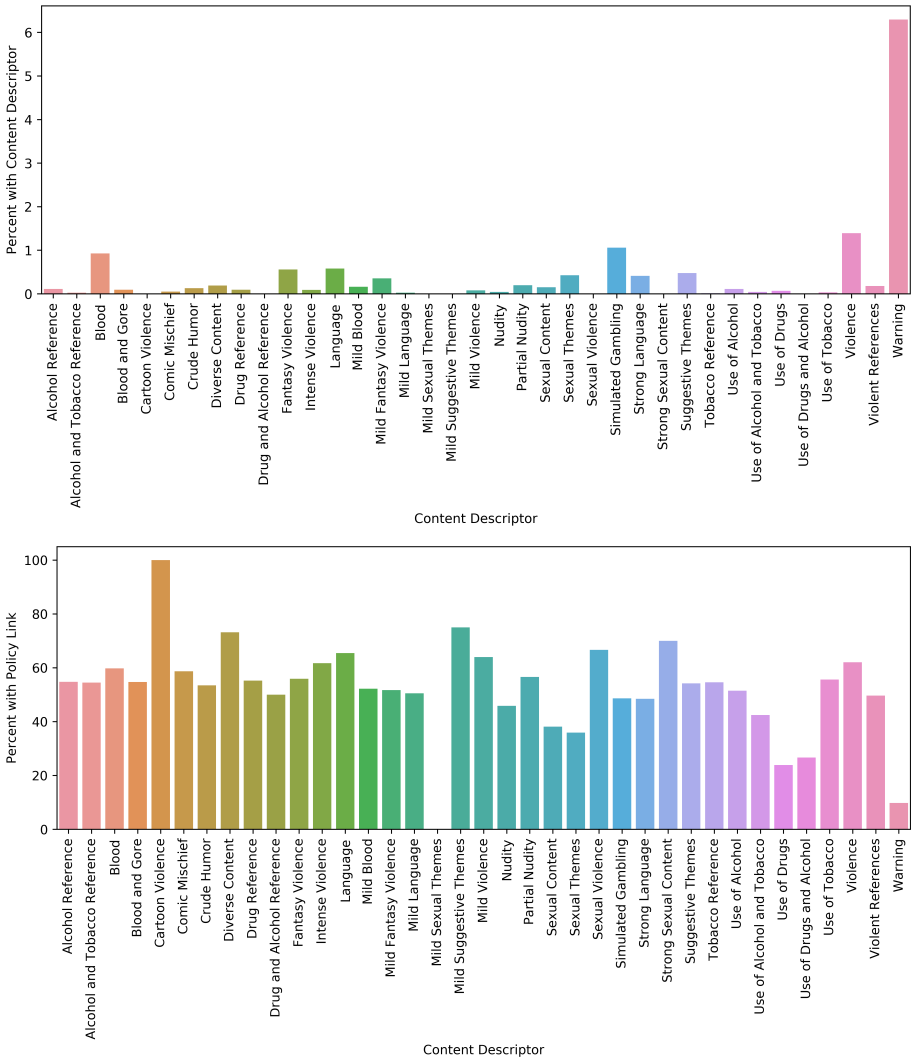

Figure 9. Percent of apps per ESRB content descriptor (top) and respective policy percent (bottom).

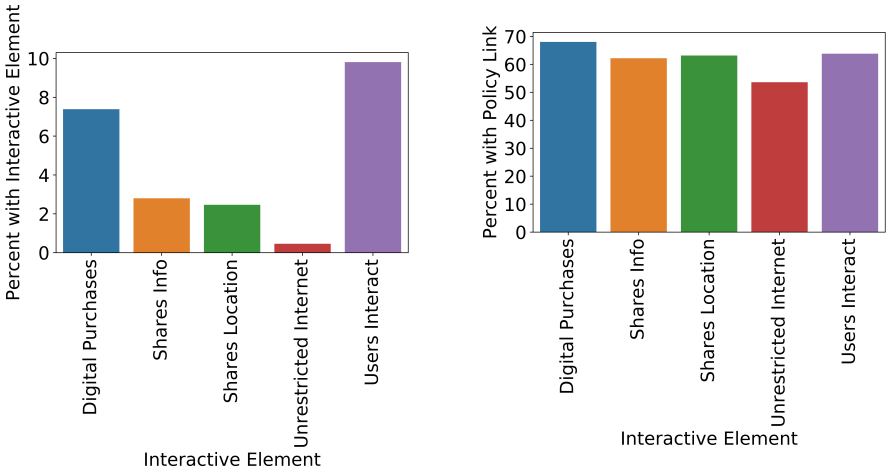

Figure 10. Percent of apps per interactive element (left) and respective policy percent (right).

5 Results

In this section we discuss the results of our logistic regression model (§ 5.2) as well as some preprocessing steps that affect those results (§ 5.1). We begin with the latter.

5.1 Preprocessing

We performed various data preprocessing steps to improve the performance of our model. Initially, we removed metadata for 8,997 apps as the data were incomplete: the metadata for those apps were missing install ranges. It may be that the Play Store page rendered by Google is sometimes incomplete.

Further, as an app may legitimately not have a rating value given a rating count of zero, we imputed missing rating values with all apps' mean rating value of 4.206.

During the time of our First and Second Crawls, the Play Store represented the number of installs per app in numerical ranges (1–5 installs, 5–10 installs, etc.). We trained models using a categorical representation for these ranges. However, we realized that the coefficients corresponding to the apps with very high install ranges were automatically eliminated by the model during training. This elimination could be the result of the relatively small count of apps with very high install ranges. The automatic removal is problematic because we want the install ranges of such apps to be taken into account when making predictions with our model. To mitigate this problem, we transformed the ranges into a quantitative variable consisting of the ranges' mean values. For example, the 10 - 50 Installs category became $(50 + 10)/2 = 30$. This quantitative variable

was retained by the model. As desired, the install ranges of apps with very high ranges are able to influence the predictions of the model.

We represent each app's publication date as the count of seconds after the currently oldest app on the Play Store was published (October 22, 2008).

5.2 Analysis

Model Description We designed our model based on scikit-learn (version 0.19.1) [27] using 67% of our data for training and 33% as a held out test set.

Our model achieves the following performance on the test set: accuracy = 67.7%, precision = 65.1%, recall = 61.5%, F1 = 63.2%. The accuracy of our model compares favorably to a baseline of always predicting that an app has a privacy policy, which would lead to an accuracy of 54.8%. We chose to use the SGDClassifier [7] instead of the standard LogisticRegression classifier [6] because stochastic gradient descent was orders of magnitude faster due to the size of our dataset [10]. We trained the SGDClassifier with the log loss function (`loss='log'`), the L1 penalty (`penalty='l1'`), 1,000 maximum iterations over the training data (`max_iter=1000`), and a stopping tolerance of 0.001 (`tol=0.001`). We choose the L1 penalty in order to get a sparse set of coefficients. As recommended [9], we ran parameter selection over the `alpha` parameter. We left the other parameters as the defaults.

We squared and cubed all the quantitative variables (date published, install range, price, rating count, and rating value). Without these transformations, our test accuracy would have been 67.2% instead of 67.7%. Although, the change in accuracy is incremental, the transformations improved interpretability; more variables were eliminated from the model. We chose not to perform log transformations because it would have made the interpretation more complicated: log(0) is undefined, and date published, install range, price, and rating count can assume zero values.

Note that the SGDClassifier requires that quantitative features be centered and scaled. We used the StandardScaler [8] and performed scaling with

$$\frac{x - \bar{x}}{s} \tag{1}$$

where x is a sample value, $\bar{x}$ is the mean of the training data, and s is the standard deviation of the training data. The values for $\bar{x}$ and s in our training data are displayed in Table 1.

Interpretation of Results Next, we explain how to interpret our model using the coefficients displayed in Table 2. Features that were eliminated from the model are not included in the table; these features neither increased nor decreased the odds of accurately predicting whether an app has a privacy policy. Note that scaling—as explained previously and shown in the Table 1—must be performed before making predictions using the model. Such scaling is benefi-

Scaling $\bar{x}$	**Scaling s**	**Feature Name**
2.480e+08	4.170e+07	date_published_relative
6.323e+16	1.816e+16	date_published_relativeˆ2
1.643e+25	6.253e+24	date_published_relativeˆ3
3.406e+05	1.541e+07	install_range
1.954e-01	3.626e+00	price
1.319e+01	1.151e+03	priceˆ2
3.117e+03	1.598e+05	rating_count
4.205e+00	6.427e-01	rating_value
7.914e+01	2.887e+01	rating_valueˆ3

Table 1. Parameters for scaling the logistic regression model's quantitative features. $\bar{x}$ and s are the mean and standard deviation of the training data, respectively. Features eliminated by the model are omitted.

Coefficient	**Odds Mult.**	**Feature Name**
-0.329		Intercept
-1.664e+00	÷ 5.280e+00	date_published_relativeˆ2
-7.437e-01	÷ 2.104e+00	category_BOOKS_AND_REFERENCE
-5.399e-01	÷ 1.716e+00	content_rating_UNRATED
-5.399e-01	÷ 1.716e+00	content_descriptor_Warning
-4.856e-01	÷ 1.625e+00	content_descriptor_Use of Drugs
-4.270e-01	÷ 1.533e+00	content_descriptor_Sexual Themes
-3.989e-01	÷ 1.490e+00	category_GAME_ARCADE
-3.501e-01	÷ 1.419e+00	category_LIBRARIES_AND_DEMO
-3.355e-01	÷ 1.399e+00	country_belarus
-3.299e-01	÷ 1.391e+00	category_GAME_ACTION
-2.964e-01	÷ 1.345e+00	content_descriptor_Sexual Content
-2.574e-01	÷ 1.293e+00	country_ukraine
-2.522e-01	÷ 1.287e+00	category_EDUCATION
-2.367e-01	÷ 1.267e+00	category_GAME_PUZZLE
-2.337e-01	÷ 1.263e+00	priceˆ2
-2.204e-01	÷ 1.247e+00	country_russia
-2.125e-01	÷ 1.237e+00	category_COMICS
-1.950e-01	÷ 1.215e+00	category_ENTERTAINMENT
-1.855e-01	÷ 1.204e+00	category_PERSONALIZATION
-1.805e-01	÷ 1.198e+00	category_GAME_BOARD
-1.682e-01	÷ 1.183e+00	country_vietnam
-1.644e-01	÷ 1.179e+00	rating_valueˆ3
-1.598e-01	÷ 1.173e+00	content_descriptor_Simulated Gambling
-1.324e-01	÷ 1.142e+00	category_GAME_ADVENTURE
-1.116e-01	÷ 1.118e+00	content_descriptor_Blood
-6.541e-02	÷ 1.068e+00	category_GAME_RACING
-2.583e-02	÷ 1.026e+00	category_MUSIC_AND_AUDIO
-2.540e-02	÷ 1.026e+00	category_ART_AND_DESIGN
-1.856e-02	÷ 1.019e+00	category_SOCIAL
-1.527e-02	÷ 1.015e+00	country_egypt
4.896e-02	× 1.050e+00	price
6.041e-02	× 1.062e+00	country_ireland
6.222e-02	× 1.064e+00	content_descriptor_Intense Violence
7.402e-02	× 1.077e+00	category_HEALTH_AND_FITNESS
7.622e-02	× 1.079e+00	category_MEDICAL
7.969e-02	× 1.083e+00	category_LIFESTYLE
8.973e-02	× 1.094e+00	category_GAME_CASUAL
9.258e-02	× 1.097e+00	rating_value
1.044e-01	× 1.110e+00	content_descriptor_Mild Fantasy Violence
1.062e-01	× 1.112e+00	date_published_relative
1.171e-01	× 1.124e+00	country_france
1.369e-01	× 1.147e+00	country_Other
1.379e-01	× 1.148e+00	category_SPORTS
1.404e-01	× 1.151e+00	category_VIDEO_PLAYERS

Coefficient	**Odds Mult.**	**Feature Name**
1.618e-01	× 1.176e+00	category_GAME_SIMULATION
1.810e-01	× 1.198e+00	interactive_element_Shares Info
2.147e-01	× 1.240e+00	interactive_element_Shares Location
2.154e-01	× 1.240e+00	install_range
2.199e-01	× 1.246e+00	country_pakistan
2.268e-01	× 1.255e+00	category_PRODUCTIVITY
2.547e-01	× 1.290e+00	country_canada
2.551e-01	× 1.291e+00	country_poland
2.704e-01	× 1.310e+00	country_india
2.867e-01	× 1.332e+00	country_australia
3.131e-01	× 1.368e+00	category_FINANCE
3.256e-01	× 1.385e+00	content_rating_EVERYONE_10_PLUS
3.517e-01	× 1.421e+00	country_germany
3.553e-01	× 1.427e+00	category_TRAVEL_AND_LOCAL
3.822e-01	× 1.465e+00	country_spain
3.919e-01	× 1.480e+00	category_GAME_CASINO
3.986e-01	× 1.490e+00	category_COMMUNICATION
4.052e-01	× 1.500e+00	country_japan
4.339e-01	× 1.543e+00	country_switzerland
4.411e-01	× 1.554e+00	country_israel
4.631e-01	× 1.589e+00	country_united kingdom
4.759e-01	× 1.610e+00	country_china
5.839e-01	× 1.793e+00	interactive_element_Users Interact
5.911e-01	× 1.806e+00	content_descriptor_Diverse Content
6.262e-01	× 1.870e+00	content_descriptor_Language
6.268e-01	× 1.872e+00	country_portugal
6.290e-01	× 1.876e+00	category_BUSINESS
6.397e-01	× 1.896e+00	country_hong kong
6.578e-01	× 1.930e+00	category_PHOTOGRAPHY
6.854e-01	× 1.985e+00	interactive_element_Digital Purchases
7.229e-01	× 2.060e+00	content_descriptor_Violence
7.297e-01	× 2.074e+00	category_SHOPPING
7.354e-01	× 2.086e+00	country_netherlands
7.903e-01	× 2.204e+00	country_usa
1.556e+00	× 4.740e+00	rating_count
2.002e+00	× 7.401e+00	category_FAMILY_MUSICVIDEO
2.157e+00	× 8.645e+00	date_published_relativeˆ3
2.505e+00	× 1.224e+01	category_FAMILY_PRETEND
3.004e+00	× 2.016e+01	category_FAMILY_ACTION
6.840e+00	× 9.344e+02	category_FAMILY_CREATE
8.913e+00	× 7.427e+03	category_FAMILY_EDUCATION
1.359e+01	× 7.998e+05	category_FAMILY_BRAINGAMES

Table 2. Coefficients of the trained logistic regression model sorted by coefficient size. A negative coefficient indicates that a feature decreases the odds of an app having a privacy policy whereas a positive coefficient indicates an increase in the odds. Odds multipliers are calculated by raising e to the coefficient.

cial as the relative sizes of the coefficients can be used to roughly compare the relative importance of the different features.[12]

The goal of our interpretation is to observe how the odds of an app having a privacy policy are affected by modifying one or more features as compared to a baseline app. For interpreting the model, note the following definition of the baseline app.

- The Undefined country was selected for the baseline app because over 82% of apps have this value, and it can be interpreted as not knowing what country the app is from.
- The ESRB content rating EVERYONE was selected for the baseline app because it is the most common, with over 84% of apps having this rating.
- Before scaling, date_published_relative ranges from 0 to 287,452,800 (corresponding to October 22, 2008 and December 1, 2017, respectively). We selected October 22, 2008 as the publish date for the baseline app, since this is the publish date of the oldest app.
- Before scaling, install_range ranges from 3 to 3,000,000,000. For our baseline app we selected 3 installs, since this is the smallest possible value.
- Before scaling, prices range from $0 (free) to $400. Our baseline app is free, since most apps on the Play Store are free.
- Before scaling, rating_count ranges from 0 to 72,979,974 ratings. Our baseline app has no ratings, since this is the smallest possible value.
- Before scaling, rating_value ranges from 1 to 5. Our baseline app has a rating value of 1, since it is the smallest possible value.
- Our baseline app has no categories, interactive elements, or content descriptors. It also does not have the Editors' Choice badge.

Given this definition, the odds of the baseline app having a privacy policy are 0.412. For details about how these odds were calculated, see Appendix 8.1.

Suppose we change the country of the baseline app to Germany (country_germany). The new odds can be calculated by multiplying the baseline app's odds by the corresponding odds multiplier from Table 2. This change gives us odds of $0.412 \times 1.421 = 0.585$. Consequently, an app from Germany has greater odds of having a policy than an app from an Undefined country. For an example of changing a quantitative variable, see Appendix 8.1.

6 Discussion

Our exploratory data analysis, logistic regression model, and longitudinal analysis (per below) may be helpful to government regulators, privacy organizations, app store operators, and others interested in understanding the state of privacy in the app ecosystem.

[12] Inherently, scaling has the disadvantage that the intercept cannot easily be interpreted because it is the y-intercept of the scaled variables.

6.1 Exploratory Data Analysis

We believe that our analysis of the features associated with apps having privacy policies (§ 4) can help regulators prioritize enforcement actions. For example, Figure 10 shows that only 63.1% of apps which describe themselves as sharing their users' locations link to privacy policies. Although, this percent is higher than the percent for the Play Store as a whole, ideally all apps which share users' locations would have privacy policies. While this finding requires further investigation, it suggests that a number of apps might not be compliant with the General Data Protection Regulation.

6.2 Logistic Regression Model

Our logistic regression model (§ 5) yields additional insights. The coefficients of the model, as displayed in Table 2, provide insight into how different features affect the odds of apps having privacy policies. Since the quantitative variables are scaled, the relative sizes of the coefficients can be used to roughly compare the relative importance of the different features. A negative coefficient indicates that a feature decreases the odds of apps having a privacy policy whereas a positive coefficient indicates an increase in the odds. For example, knowing that an app is from the Books and Reference category divides the odds by 2.104 (that is, decreases the odds by approximately 50%), whereas knowing that the app offers in-app purchases multiplies the odds by 1.985 (that is, increases the odds by approximately 100%). Our model eliminated redundant and uninfluential features, which would otherwise serve as noise and obscure the truly meaningful features.

In addition to identifying features that affect the odds of apps having policies our model lends itself to another use case: the model can identify the apps with the highest probability of having policies but which in actuality lack such (that is, false positives). The fact that these apps do not have policies makes them stand out from similar apps. For example, there are many apps with more than 100,000 ratings and millions of installs, which our model predicts would have policies but which actually lack them. As some of these apps are from major companies, the policy absence strikes us as an oversight instead of a lack of knowledge about applicable privacy regulation. In those instances, regulators might find it worthwhile to simply notify the affected companies of their shortcomings to mitigate potential non-compliance with privacy laws and regulations.

6.3 Longitudinal Analysis

Another interesting finding comes from comparing our three crawls. In our First Crawl (August 28 through September 2, 2017), 41.7% of apps had privacy policies. This number increased to 45.2% in the Second Crawl (November 29 through December 2, 2017), and to 51.8% by the time of the Third Crawl (May 11 through May 15, 2018). Further, the number of apps discovered by our crawling techniques decreased over the course of the crawls: $n = 1,423,450$ apps (First

Crawl), $n = 1,163,622$ apps (Second Crawl), and $n = 1,044,752$ apps (Third Crawl). Notably, only about 179K apps were originally discovered in the Third Crawl; we seeded the database with the app identifiers collected by the Second Crawl in order to gather the metadata of more apps. One possible explanation for these changes could be Google's curation of apps on the Play Store in between our crawls. After all, Google announced removing apps that collect "Personal and Sensitive Information" but do not have privacy policies [5]. Another possibility could lie in Google's changes to "limit visibility" of certain apps, preventing us from discovering them in our crawl even if they are still present in some form on the Play Store [5]. The sharp decrease in the number of apps discovered by our recursive crawling technique in the Third Crawl shows that Google changed how they recommend related apps. If the increase in the percent of apps with policy links was caused by Google's curation of the Play Store, our findings would show how action by ecosystem managers can have a substantial effect in potentially increasing privacy compliance. Regardless of the explanation, the increase is a step in the right direction as it certainly does not decrease privacy.

7 Conclusion and Future Work

In this study we discussed our exploratory analysis of features associated with apps having privacy policies (§ 4), presented our logistic regression model for predicting whether apps actually have privacy policies (§ 5), and explained how our work might be useful to government regulators as well as other organizations and individuals interested in privacy (§ 6).

Our exploratory analyses yielded novel insights (§ 4). Most notably, we discovered that only 63.1% of apps which are described as sharing their users' locations link to privacy policies. By analyzing the metadata of over a million apps, we are able to make conclusions about the privacy landscape in the Android app ecosystem. In our repeated crawls of the Play Store, we discovered possible evidence of Google's actions contributing to an increase in the percent of apps with privacy policies (§ 6.3). The coefficients of our logistic regression model show how individual features affect the odds of apps having privacy policies (§ 5). The model can also be used to identify apps which stand out from similar apps for not having privacy policies (§ 6.2).

A number of areas for future work remain. First, this study focused on the US Google Play Store. It would be worthwhile to perform similar analyses on Play Stores localized for European countries. This would give us insight into how different data protection frameworks affect the prevalence of privacy policies. A longitudinal analysis might even give insight into the effects of new legislation on the privacy landscape. We would welcome the opportunity to engage with European researchers, regulators, privacy organizations, and other parties to perform such comparative analyses.

Second, in the course of conducting our study we observed several examples of privacy policies or parts thereof appearing across seemingly unrelated organizations. In some cases, this repetition of policy language seems to indicate

the use of privacy policy generators. However, it sometimes appears that language was simply copied from one policy to another. It would be worthwhile to systematically examine privacy policy reuse across the entire Play Store and beyond. Based on previous work showing a generally positive relationship between company size and whether companies have privacy policies, we hypothesize that smaller organizations may be more likely to reuse policy text [2].

Third, we are working on a large-scale system for comparing the actual practices of apps with the practices described in their privacy policies—using static code analysis and natural language processing, respectively. By analyzing apps' code and privacy policies, our system will automatically flag discrepancies between the two. The system will be a substantial advancement over the work described in this study, because simply knowing whether an app has a privacy policy or not is typically insufficient to determine non-compliance with regulation. In particular, apps that do not collect or share any personally identifiable information are generally not required to have a privacy policy. Also, simply having a privacy policy is insufficient to guarantee compliance, because the privacy policy may not describe all of an app's practices. However, we view our metadata analysis as complementary to this more in-depth analysis; our metadata analysis can help prioritize investigation of the discrepancies flagged by our in-depth analysis.

This study is our first large-scale analysis of the privacy landscape of the Play Store. However, there are still many untapped research opportunities in this area. We plan to use the infrastructure we developed for this analysis for additional large-scale analyses in the future.

8 Appendix

8.1 Odds Calculations

Here we provide additional details about how the baseline app's odds were calculated, and how to interpret the model's quantitative variables.

Logistic regression models operate directly in terms of log(odds). For interpretability, log(odds) are easily converted to odds:

$$e^{\log(\text{odds})} = \text{odds} \tag{2}$$

Under the definition of the baseline app in § 4, the log(odds) of the baseline app having a privacy policy can be calculated by substituting the coefficients of Table 2 into the following equation:

$$\log(\text{odds}(\text{policy} = \text{True})) =$$
$$b_0 + b_{\text{date\_published\_relative}} * \text{date\_published\_relative\_scaled} + \ldots \tag{3}$$

where b_0 is the intercept, $b_{\text{date\_published\_relative}}$ is a feature coefficient, and date_published_relative_scaled is a scaled feature value. Note that the full equation would include terms for all of the coefficients in Table 2. From this equation,

we calculate the log(odds) $= -0.887$ and odds $= e^{-0.887} = 0.412$ of the baseline app having a privacy policy.

Next, we give an example of changing a quantitative variable. Suppose we start with the baseline app, which has no ratings, and increase the rating_count to 1,000,000. First, we scale rating_count[13] using the coefficients from Table 1:

$$\Delta\text{rating\_count\_scaled} = \frac{1,000,000 - \text{rating\_count\_baseline}}{s_{\text{rating\_count}}} \approx \frac{1,000,000}{1.598 * 10^5} \approx 6.258 \tag{4}$$

Next, we simply multiply this scaled value by its corresponding coefficient from Table 2 and add it to the log(odds) of the baseline app. This gives us log(odds) $=$ 8.850, or equivalently odds $= 6,974$. According to our model, an app with 1,000,000 ratings has much greater odds of having a privacy policy than an app with no ratings.

References

1. Almuhimedi, H., Schaub, F., Sadeh, N., Adjerid, I.: Your location has been shared 5,398 times!: A field study on mobile app privacy nudging. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (2015). https://doi.org/10.1145/2702123.2702210, http://dl.acm.org/citation.cfm?id=2702210
2. Balebako, R., Marsh, A., Lin, J., Hong, J.I., Cranor, L.F.: The privacy and security behaviors of smartphone app developers. Workshop on Usable Security (2014), http://repository.cmu.edu/hcii/265/
3. Blenner, S.R., Kollmer, M., Rouse, A.J., Daneshvar, N., Williams, C., Andrews, L.B.: Privacy Policies of Android Diabetes Apps and Sharing of Health Information. JAMA **315**(10), 1051–2 (Mar 2016). https://doi.org/10.1001/jama.2015.19426, http://jama.jamanetwork.com/article.aspx?doi=10.1001/jama.2015.19426
4. Bouchard, B., Suzuki, K.: Find great apps and games on google play with the editors' choice update. https://www.blog.google/products/google-play/find-great-apps-and-games-google-play-editors-choice-update/ (Jul 2017), accessed: May 20, 2018
5. Clark, B.: Millions of apps could soon be purged from google play store. https://thenextweb.com/google/2017/02/08/millions-apps-soon-purged-google-play-store/ (Feb 2017), accessed: May 20, 2018
6. scikit-learn developers: sklearn.linear_model.logisticregression. http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html, accessed: May 20, 2018
7. scikit-learn developers: sklearn.linear_model.sgdclassifier. http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html, accessed: May 20, 2018
8. scikit-learn developers: sklearn.preprocessing.standardscaler. http://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html, accessed: May 20, 2018

[13] We can ignore rating_countˆ2 and rating_countˆ3 because they were eliminated from the model.

9. scikit-learn developers: Stochastic gradient descent: Tips on practical use. http://scikit-learn.org/stable/modules/sgd.html#tips-on-practical-use, accessed: May 20, 2018
10. scikit-learn developers: Choosing the right estimator. http://scikit-learn.org/stable/tutorial/machine_learning_map/index.html (2017), accessed: May 20, 2018
11. d'Heureuse, N., Huici, F., Arumaithurai, M., Ahmed, M., Papagiannaki, K., Niccolini, S.: What's app?: a wide-scale measurement study of smart phone markets. dl.acm.org https://dl.acm.org/citation.cfm?id=2396759
12. Entertainment Software Rating Board (ESRB): Esrb ratings guide. https://www.esrb.org/ratings/ratings_guide.aspx (2015), accessed: May 20, 2018
13. Fahey, K.: Recognizing android excellence on google play. https://android-developers.googleblog.com/2017/06/recognizing-android-excellence-on.html (Jun 2017), accessed: May 20, 2018
14. Federal Trade Commission: Mobile privacy disclosures. https://www.ftc.gov/os/2013/02/130201mobileprivacyreport.pdf (Feb 2013), accessed: May 20, 2018
15. Federal Trade Commission: Children's Online Privacy Protection Rule ("COPPA"). https://www.ftc.gov/enforcement/rules/rulemaking-regulatory-reform-proceedings/childrens-online-privacy-protection-rule (Aug 2015), accessed: May 20, 2018
16. FTC: Privacy online: A report to congress. https://www.ftc.gov/reports/privacy-online-report-congress (Jun 1998), accessed: May 20, 2018
17. Google: Designed for families. https://developer.android.com/distribute/google-play/families.html, accessed: May 20, 2018
18. Google: Content ratings for apps & games. https://support.google.com/googleplay/android-developer/answer/188189?hl=en (2017), accessed: May 20, 2018
19. Google: Ratings questionnaire help. https://support.google.com/googleplay/android-developer/topic/6169305?hl=en&ref_topic=6159951 (2017), accessed: May 20, 2018
20. Google: Requesting permissions. https://developer.android.com/guide/topics/permissions/requesting.html (2017), accessed: May 20, 2018
21. California Department of Justice: Attorney General Kamala D. Harris secures global agreement to strengthen privacy protections for users of mobile applications. http://www.oag.ca.gov/news/press-releases/attorney-general-kamala-d-harris-secures-global-agreement-strengthen-privacy (Feb 2012), accessed: May 20, 2018
22. California Department of Justice: Making your privacy practices public. https://oag.ca.gov/sites/all/files/agweb/pdfs/cybersecurity/making_your_privacy_practices_public.pdf (May 2014), accessed: May 20, 2018
23. Kelley, P.G., Cranor, L.F., Sadeh, N.: Privacy as part of the app decision-making process. CHI p. 3393 (2013). https://doi.org/10.1145/2470654.2466466, http://dl.acm.org/citation.cfm?doid=2470654.2466466
24. Lin, J., Liu, B., Sadeh, N., Hong, J.I.: Modeling Users' Mobile App Privacy Preferences - Restoring Usability in a Sea of Permission Settings. Proceedings of the Twelfth Symposium on Usable Privacy and Security (2014), http://dblp.org/rec/conf/soups/LinLSH14
25. Lin, J., Sadeh, N., Amini, S., Lindqvist, J., Hong, J.I., Zhang, J.: Expectation and purpose - understanding users' mental models of mobile app privacy through crowdsourcing. UbiComp p. 501 (2012). https://doi.org/10.1145/2370216.2370290, http://dl.acm.org/citation.cfm?doid=2370216.2370290

26. Palmer, J.: After several years of service, the google play top developer program is being put to rest. http://www.androidpolice.com/2017/05/05/several-years-service-google-play-top-developer-program-put-rest/ (May 2017), accessed: May 20, 2018
27. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
28. Sadeh, N., Acquisti, A., Breaux, T.D., Cranor, L.F., McDonald, A.M., Reidenberg, J.R., Smith, N.A., Liu, F., Russell, N.C., Schaub, F., Wilson, S.: The Usable Privacy Policy Project: Combining Crowdsourcing, Machine Learning and Natural Language Processing to Semi-Automatically Answer Those Privacy Questions Users Care About. Carnegie Mellon University Technical Report CMU-ISR-13-119 pp. 1–24 (Dec 2013), http://reports-archive.adm.cs.cmu.edu/anon/isr2013/CMU-ISR-13-119.pdf
29. Statista: Number of apps available in leading app stores as of march 2017. https://www.statista.com/statistics/276623/number-of-apps-available-in-leading-app-stores/ (2017), accessed: May 20, 2018
30. Sunyaev, A., Dehling, T., Taylor, P.L., Mandl, K.D.: Availability and quality of mobile health app privacy policies. Journal of the American Medical Informatics Association **16**(4), 16–6 (Aug 2014). https://doi.org/10.1136/amiajnl-2013-002605, https://academic.oup.com/jamia/article-lookup/doi/10.1136/amiajnl-2013-002605
31. Viennot, N., Garcia, E., Nieh, J.: A measurement study of google play. In: The 2014 ACM international conference. pp. 221–233. ACM Press, New York, New York, USA (2014). https://doi.org/10.1145/2591971.2592003, http://dl.acm.org/citation.cfm?doid=2591971.2592003
32. Wang, H., Liu, Z., Guo, Y., Chen, X., Zhang, M., Xu, G., Hong, J.: An Explorative Study of the Mobile App Ecosystem from App Developers' Perspective. In: the 26th International Conference. pp. 163–172. ACM Press, New York, New York, USA (2017). https://doi.org/10.1145/3038912.3052712, http://dl.acm.org/citation.cfm?doid=3038912.3052712
33. Zimmeck, S., Wang, Z., Zou, L., Iyengar, R., Liu, B., Schaub, F., Wilson, S., Sadeh, N., Bellovin, S.M., Reidenberg, J.: Automated analysis of privacy requirements for mobile apps. In: 24th Network & Distributed System Security Symposium (NDSS 2017). NDSS 2017, Internet Society, San Diego, CA (February 2017)