A Bayesian Evaluation of User App Choice in the Presence of Risk Communication on Android Devices

Behnood Momenzadeh
Indiana University
smomenza@iu.edu

L Jean Camp
Indiana University
ljcamp@indiana.edu

ABSTRACT

In this work we empirically explore the possibility that people lack the information to make risk-aware decisions when choosing between mobile apps, and if given such information would change their behavior. Specifically we examine the choice of apps by users when risk information is embedded in the display of apps. Currently, no such information is readily available. Despite the presence of permissions information, it is not cognitively feasible to compare apps on permission, nor security or privacy in current app stores. One component to resolving this lack of information is the creation of clear, effective risk communication at time of app selection. One core test of risk communication is if it influences decision-making. Here we test indicators that allow users to differentiate the risk associated with apps, and examine the impact on decision-making in four app categories. We use an experimental model grounded in medical interventions, where we add an intervention in multiple situations (in this case app categories) and compare these to the pre-existing baseline. The question we address here is not if such an indicator can be reliably generated, but rather if were clearly indicated would it make a difference? To answer this we built an extended Android Play Store that embedded indicators using the lock icon as a cue. We recruited sixty participants to test the interaction using tablets running the extended store on Jelly Bean. The Play Store was otherwise unaltered, and included the standard user ratings, download count, and permissions interface. The result was that participants systematically choose apps with lower ratings or lesser download counts instead choosing apps with higher ratings with respect to risk. We compare our results to the users' behavior in Android Market, indicating that individuals not only prefer higher privacy with no loss of functionality, but also that some participants may trade-off functionality for privacy.

KEYWORDS

Risk Communication, Android Security, Android Play Store, Android Security Rating

ACM Reference Format:
Behnood Momenzadeh and L Jean Camp. 2019. A Bayesian Evaluation of User App Choice in the Presence of Risk Communication on Android Devices. In *Proceedings of , , ,* 10 pages.

, ,

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

There is evidence that privacy and security are both subject to information asymmetry, in other words these are lemons markets [4]. In the smart phone domain, a permissions model is used to provide information to support informed choice and address this information asymmetry. These permissions allow access to phone functionality and user data. Yet, significant research has shown that people do not understand permissions. If the app market is a lemons market with respect to security and privacy, this can explain part of this privacy paradox where individuals cannot distinguish apps based on their permissions and the associated information risks. If the current interaction model is inadequate to support permissions-aware decision making, then the existence of signals can make a difference in selection of apps.

Choosing to download or use an app can be a simple decision of evaluating its costs and benefits. The benefits can be said to include security and privacy. On the other hand, information exfiltration is part of the cost, for instance through paying via personal data. With the current interaction design it is difficult for an individual to evaluate the security costs and benefits. It is possible to compare apps based on privacy and security. Yet this requires 1) an understanding of the permissions model, 2) an understanding of the risk associated with specific permissions, and 3) the cognitive work to compare the various options provided in the permissions manifests. One way to address this is to provide an easy to understand indicator to distinguish between low risk and high risk apps. We refer to such indicator as a signal.

To explore the efficacy of such a signal, we developed an Android Play Store interaction that included risk indicators which are visible in the listing of apps. We illustrated that this changed the choices about which apps to install with users handling Android Nexus 7 tablets. With an interaction that provides information about the permissions aggregated into a single indicator, individuals choose apps having better safety rating over those with more downloads or higher app ratings.

2 RELATED WORK

There are strong arguments that people do not understand permissions and may change their behavior if provided clear communication. There is also research showing that people accept over-privileging, and that individual concerns about information risks do not result in changed behaviors. In this experiment, we test these conflicting arguments by building on the underlying concept of information asymmetry and providing economic signals. An economic signal distinguishes between two otherwise indistinguishable choices. Distinguishing between apps today requires

an understanding of the app permissions model, including their privacy implications.

Currently, risk communication is grounded in access control using a permissions model. In 2011, researchers at Berkeley found that even developers do not understand the permissions model [16]. Systematic over-privileging resulted from confusion in permission naming and permission inheritance. Developers also bundle permissions, requesting entire classes when only one permission is needed. In the case of WiFi permissions, code reuse and popular but confusing documentation seem to create confusion and risk. Apps were found to include systems permissions that are refused in practice by the OS.

Given the complexity of permissions for developers, expecting users to understand them seems optimistic. Rajivan et al. also found a widespread lack of understanding measured by the ability to correctly identify the function of a permission using multiple choice [34]. A common statistic (from Felt's team in 2012) was that only 17% of people look at permissions [17]. 17% is the highest percentage reported in behavioral research since then. For example, despite 78% of participants indicating that they examine permissions before installing an app, Rajivan et al. found that fewer than 8% actually viewed a permission [34].

Multiple approaches have been proposed to increase user comprehension of permissions to improve decision-making. Thus far, however, the predominate social effect on app downloads appears to be the importance of download counts as a source for decision support [24]. Such indicators may be quite misleading, as Morton found, in both focus groups and a large scale survey. Participants in both experiments reported by Morton, indicated concerns about privacy and the conviction that widespread adoption leads to acceptable privacy and security settings [32].

There is an argument that people will change their behavior when given simple indicators of security/privacy. Forget led an investigation into cognitive engagement on mobile devices in 2016. The team concluded that people need concise, precise, and simple to use security interactions for these to be effective because of low levels of user engagement [18]. Using the socially-based icon of eyeballs, individuals were shown to change privacy behavior when the feedback was based only on app access to location information. [36]. When the eyes were used to indicate general permissions, the result was quite inconsistent [8]. Later Liccardi et al. used eyes to communicate a risk score for apps, and to draw attention to risky permissions [28]. Again there were significant results, but not consistent significant results across all categories.

Providing aggregate information about permissions requirements increases users expressions of their awareness of permissions and associated risks [27]. Vila et al. as well as Milne et al. also argue that privacy is a lemons market on the web where additional clarifying information can change decision-making [31, 41]. Tsai and colleagues found that clear information about privacy on a website resulted in consumers willingness to pay a premium for privacy [40]. They found a willingness to pay for items that are widely considered privacy-sensitive (a sex toy) and items that are not considered privacy-sensitive (batteries, ones that were not compatible or related to the toy). Changing a privacy interaction on the web changed willingness to share information [1, 25].

Yet the counter argument can be made, that perhaps people are reasonably unconcerned. It is possible that people will not respond to privacy signals. Certainly individuals ignore warnings on desktops [9, 13], so there is no certainty that they would engage with comparable icons on a mobile phone. For example, the results of an early investigation of willingness to pay to hide or to prevent exposure of information is illustrated in the title: "When 25¢ is too much" [20]. While there were some willing to pay a premium to hide information, most valued information concealment on the order of pennies. In later work in the mobile domain, a plurality of users stated a willingness to accept risk to obtain either the desired functionality, a free app, or a combination [21].

Consider the cognitive work necessary to compare the various options provided in the permissions manifests as mentioned above. In 2015, Acquisti et al. showed that people are cognitive misers and this applies in security and privacy [2]. This is particularly important in mobile computing because individuals may respond differently to actual devices as opposed to a simulated device. Tasks on tablets and mobile devices are quite likely to have different cognitive responses than desktop computers. The same task on a desktop requires more cognitive effort than on a mobile phone. The simple proximity of an item to a hand increases the likelihood of response to a stimuli. Such proximity also reduces the speed of evaluation while increasing the comprehension [43]. Both distance from the face and the choice of single or double hands changes cognitive response [12], with two hands requiring a lessor interruption for the same focus. Since a significant component of our experiment embeds comprehension, we choose the more labor-intensive method of building a functional app store and recruiting individuals for actual use of the app over the option of a larger sample size in MTurk.

Our requirements for this research was a consistent, repeatable rating that could be implemented by any other group wishing to repeat or extend the research. There has been numerous research projects to quantify the risk of an app. We can categorize the rating systems into 4 different systems as described below:

Manifest-based: These projects focus on the permissions an app requests with regard to the category of its app. For instance, if a flashlight app asks for access to contacts, it looks as a malicious activity for these rating systems [7, 38].

Static analysis: In this category, projects focus on analyzing the source code of an app to identify suspicious accesses to phone resources [15, 37].

Dynamic analysis: These projects examine apps during run time, particularly focusing on potential misuse of resources [6, 11, 14].

User intent: In these projects, researchers try to infer the intent of the user when he/she uses an app and grants a permission [29, 35, 42].

The purpose of building on permissions as a basis for our ratings was three fold. First, we wanted to use information that is already provided in the marketplace as opposed to embedding a new standard or ratings mechanism for risk. Second, permissions provide information about security and privacy, whereas a privacy-only or security-only rating would have to be generated by the researchers. It is reasonable to assume that the mobile team at Android has a better understanding of risk concerns on their own platform than

we could develop for this experiment. Third, and most importantly, we seek to make a contribution about supporting decision-making in the marketplace. It is not our intention to define the best possible rating system. The ratings used in this experiment are taken from Privacygrade project. However, any source of privacy or security ratings could be used if it is in fact true that these indicators impinge individual decision-making.

We make no claims about the ratings except internal consistency. The goal of this research is to evaluate if indicators found to be significant in previous research would result in changes in app selection in a realistic Play Store environment. We make no arguments here about this being the optimal rating, as the relationship between privacy and security risks as well as privacy and security perceptions of these risk make designing such a rating a difficult challenge. Our ratings are an internally consistent safety rating. More locks imply a less risky, more secure and/or more private app compared to fewer locks.

For the visualization, we choose locks based on previous work. In addition to examination of specific cues described above Rajivan et al. implemented pairwise examinations of different risk icons and padlocks were found the most effective in changing risk-based decision-making [34].

3 EXPERIMENT DESIGN

Do people change their app selections when provided information that distinguishes apps based on their permission requests that allows risk comparisons while choosing an app? If there is significant change in behavior this supports an argument that individuals will act on permissions if the information is presented in clear, simple, and timely. Whether the change in behavior is a result of improved usability or risk communication, the availability of information in comprehensible, usable, and transparent form is essential to a functioning market. Risk communication is most effective when risk mitigation is integrated and immediately actionable. Providing information while the user is comparing apps makes risk avoidance cognitively simple.

Our goal was to compare behavior of our participants with the behavior of participants in the overall marketplace. Thus we altered the store interaction, and asked individuals to select multiple apps. We rank the apps in the experiment based on the number of downloads in our experiment and compare them against the rank of them in the actual Play Store based on their number of downloads. We selected four categories of apps. We selected apps where the functionality is almost identical and the use of information is very consistent across the options (flashlights). We selected apps where there is great variance in information exfiltration, but still little difference in functionality (weather), and apps where there is some variance in information exfiltration and functionality (photos). The fourth category of apps varied greatly in both functionality and use of information (games). We had individuals rank the top four apps by selecting these to install on a tablet provided during the experiment.

Because of the risks associated with information exfiltration, the rankings include components of privacy and security. For example, if a permission to view the photos is used only with personal photos, then this is a privacy issue. However access to photos if the camera is used to record payment or identifying then this is a security issue, as shown by [26]; as well as the practice of preventing mobile phones from military and other sensitive locations. Sound recordings can be a privacy violation or a security vulnerability [36]. In the example of photos, evaluating if the photos are manipulated on the cloud or on the local phone is itself a research challenge [33]. Additionally, using preferences within a category inherently mitigates operational concerns, where permissions that create privacy and security risks are inherent to the category; for example, photo apps require access to a camera while weather may not.

Our goal was to determine if people change their choices in the face of a range of privacy options when these are communicated. Our intent in experiment design was to answer the following research questions.

RQ1: In the absence of differing risk indicators, is our group of participants indistinguishable from the Android market as a whole?

RQ2: When the functionality of the apps is the same but the risk differs, do our participants make choices that are indistinguishable from the Android market as a whole?

RQ3: When the functionality is different but the risk is the same, do our participants make choices that are indistinguishable from the Android market as a whole?

RQ4: When the functionality and the risk both differ, do our participants make choices that are indistinguishable from the Android market as a whole?

In terms of RQ1, for flashlight apps the functionality of the apps is the same and the use of permissions is similar. Note that we are not making a statement about the desirability of the patterns of use of permissions in flashlights, only noting that they have a leptokurtic distribution.

In terms of RQ2, in weather apps the functionality varies little, but the safety rating varies significantly.

In terms of RQ3, in games the functionality varies between apps, and the privacy rating varies slightly.

In terms of RQ4, in the category of apps classified as photos, the functionality is less self-similar than in the weather category. Conversely the range of risk ratings varies less than that of the weather category.

As in RQ1 and RQ3, the safety ratings do not differ much, if our market is a good representation of actual Play Store, the distribution of participants' choices should be similar to the distribution of users' choices in Google Play Store. In other words, as the security ratings do not vary, we are not adding information to participants and as a result we expect the distribution of their choices to follow the distribution of Google Play Store users. We use the result of these two research questions to verify our experiment design. In RQ2 and RQ4, we have varying safety ratings. If the distribution of the participants' choices differ from that of Google Play Store users, it means our safety ratings have made difference.

If there is no change in the perturbation of ranking between these four categories in our experiment versus the Play Store, this would support the contention that individuals are not concerned with app risks. (Note the risks which we indicate with the lock

icon are grounded in privacy grade, and based on crowd-sourced evaluations of permissions.)

If the results of our participants‘ selection had proven to be different under all categories then we could make no conclusions.

If there were a greater change when the functionality is the same, but the risk varies this indicates that people would mitigate risks if there is no loss in functionality.

If there is a difference between the distributions when there are differences between functionality and risks, this indicates that people would mitigate risk even if there is a corresponding loss of functionality. Recall the basis for comparison is the overall Android market. The assumption we are making is that the ordering of the market provides information about the benefit or desirability of the functionality of the app. That is, we assume that the apps when ordered by download in the Google Play Store reflect an aggregate valuation of the app.

In order to evaluate our sample against the existing market, we choose a Bayesian approach where we can compare the magnitude of the changes of the distribution, and not simply the likelihood that a given result is different. A frequentist approach, such as two sided t test, compares privacy means; with different formulations under different sets of assumptions. However a Bayesian analysis allows us to compare distributions and determine the degree and likelihood to which the distributions are different. Ensuring a robust frequentist result requires that the sample sets be representative for the results to be valid for a larger population. This is ideally implemented by sampling a representative population. Such a sampling of Android users was not feasible for an academic as opposed to an industrial research project, instead we integrated into the method the ability to compare the representativeness of the behavior of the population when they were given no actionable additional risk information. If the selections of the participants were significantly different when there was no variation in the safety ratings, then we could only conclude that our interaction perturbed behavior with no indication that there was any risk communication.

That is, we evaluated our intervention as a comparison between our population and the existing data on Android market behavior. This is analogous to comparing a medical treatment group to those receiving a known intervention with well documented outcomes. In this way, instead of comparing a segment of a potentially highly non-representative sample we can test if our sample is representative, as shown in the analysis of RQ3 and RQ1. We can then determine the level of difference in our populations, as shown in our analysis of RQ2 and RQ4 (i.e., where the risk ratings differ). To meet the goal that our population would be somewhat representative in terms of computing expertise, we did not recruit any participants from within the University itself.

As our interest is in the impact of participant decision-making, we developed an app store with the goal of being cognitively identical to the current Play Store with the only difference being the presence of signals about the risk. If there is no change in behavior then our distribution of apps selected should be the same as the distribution of apps selected in the Android Play Store. That is, we should be sampling from a well known distribution and our sample should be representative. A null hypothesis test would provide information as to if two samples could be assumed to be same. In contrast, a Bayesian approach allows us to allocate the likelihood or credibility across a range possibilities. Thus we state our research questions as inquiries for which a range of results is possible, rather than as null hypotheses.

Every app in the current Android market place has a rating out of five and apps are presented in order of popularity and match to the search. The apps were presented in the same order in our work. To calculate the privacy and security of each app, we use ratings from privacygrade, a project that falls into user intent category [29, 35] . Privacygrade assigns ratings of A to D for different apps which we convert to 5 to 2 locks. If the app is too new and therefore it has not been processed by privacygrade yet, the app will be assigned 1 lock. However, any arbitrary risk rating could be implemented as long as it is consistent across categories and the apps within these categories. We added our ratings to an otherwise identical Play Store and implemented this augmented interaction on Android Nexus 7 tablets. We recruited a diverse participant population through outreach at the public library and the Farmers Market. Our sixty participants chose apps from our alternative Play Store and are compared against Android Play Store users. With the users of Android Play Store selecting apps based on user ratings and downloads, while the experimental group‘s choices reflected to the information that was then in the Play Store in addition to our risk communication.

Our system is designed to accept arbitrary ratings as long as these can be normalized from 0 -5. Our model embeds economics not only in the conception of the problem in terms of user communication as economic signaling, but also in the experimental implementation that assumes that there might be competing sources of ratings based on the priorities of the user.

It is worth mentioning that we specifically chose to accept fewer participants in order to have a hand-held interaction, as opposed to a Play Store interaction on MTurk. We developed a modified Play Store because of findings from the psychology of decision-making that show cognitive differences in interactions between hand-held and desktop devices [12, 43]. We were also informed by the work of Amrutkar, Singh, Verma and Traynor which illustrated the differences between security interactions in the web browser on a desktop versus a hand-held device [5].

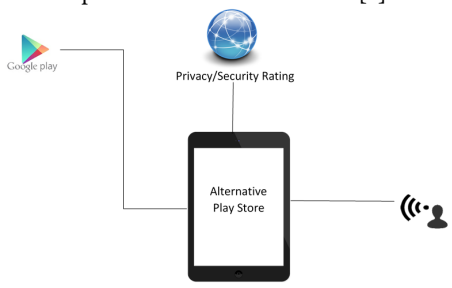

Figure 1: Our system overview

The alternative Play Store architecture is shown in Figure 1. We used Android Nexus 7 tablets with Jelly Bean. We modified a Play Store that was initially developed as part of μg [30] project. Although it is no longer maintained, we found it suitable for our needs. It uses the Google play store APIs to get the data necessary for our experiment: user ratings, number of downloads, descriptions, app display, and list of permissions.

We added the risk rating to the data queries, automated the user login, changed the GUI to include the risk rating, changed the way installation worked (which we will talk about later in methodology section). We made sure that we sent minimal requests to fetch risk ratings both not to flood the server with duplicate requests and to optimize the performance of our alternative play store. We save the rating fetched for a version of a particular app so we do not have to resend the request.

4 METHODOLOGY

A core design goal was to make the experimental interaction as close as possible to the actual experience of a user interacting with the Android Play Store marketplace. The user could search, choose, download, and install the apps he/she chose to do. But after running our first pilot experiment we saw that downloading apps was quite time consuming, resulting in fewer people willing to complete the entire experiment. In our pilot with local wireless availability the experiment could take up to 60 minutes (much longer than 15 minutes we had estimated for each participant). This was a function of recruiting participants off campus, with low bandwidth wifi in the mall, library, and town square. As a result, we decided not to download the apps and only show a "successful installation" pop up as soon as the participant clicked on the download/install button.

The essence of our experiment is asking users to select apps as if they were to install an app from that certain category and evaluate the resulting decision in terms of the risk rating (i.e. locks), download counts, and community rating (i.e. stars). As shown in Figure 3 the user could have read the permissions before installing the app using the button below the screen shots of the app.

We asked each participant to select four apps from four different categories for a total of sixteen apps. We provided the tablet, and asked participants to use specific search terms for each category. The categories were Flashlight, Photos, Games, and Weather. To make sure all the participants saw the same results we ensured that each participant used the exact same term (i.e. the name of the categories mentioned above). The experiment was subject to IRB review and approval. We did not describe the purpose of the experiment as being grounded in security. Also, we did not bring the participant attention to the indicators. Our goal was not to inform participants about risk to observe their choices without priming. We made sure the participants had prior experience with Android devices. We also changed the order of categories (flashlights- weather- photos- games) between participants to avoid biases as a result of the order of the categories.

When the keyword for each category was searched during the experiment, participants were presented with many apps to choose from. We compare installs during the experiment with downloads from Google's app store which with identical search results. We also perform an analysis to see if we have affected users' decisions.

We had experimental participants make multiple choices to create a situation where the participant is given a choice between less risk or higher ratings/popularity. That trade-off was possible in categories chosen to answer RQ2 and RQ4 categories (weather and photos). However, the flashlight category all apps provided essentially equivalent privacy. The apps in flashlight category had

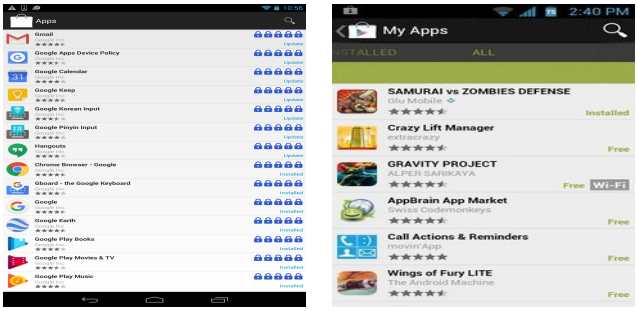

Figure 2: Our Alternative Play Store on the left Compared to Android Play Store on the right

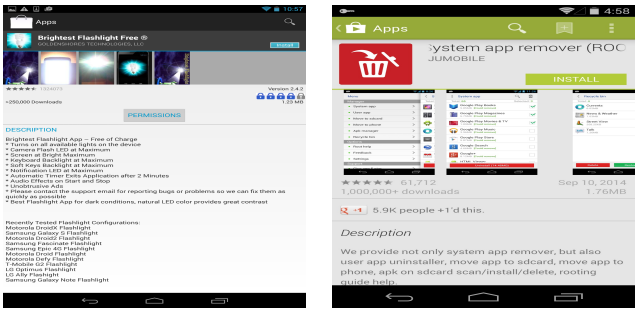

Figure 3: Our Alternative Play Store on the Left Compared to Android Play Store on the Right

high safety ratings because of the uniformity of permissions use in that category.

After the participants completed the selection of apps, we asked them to fill out two surveys to describe their experience with the marketplace. One of the surveys was a workload task survey, the NASA task load index [23]. The second survey consisted of demographic questions and also the participants' self-reported habits while installing applications from app store.

5 RESULTS

We had 58% male and 42% female participants. Across all participants 22% said they will "almost every time" or "always" check permissions while installing an app. However, only 7% of our installations were preceded with a check of an application's permissions. This is similar to previous experimental results, as noted above [34, 44]. A majority (79.6%) of the participants declared they have refused to continue with the installation of an app because of its permissions before. In our experiment, there was only one instance in which a user did not continue with the installation after viewing the permissions.

If our participants are representative samples of Android participants then the distribution of selection of apps should match the distribution of selection of apps in the Play Store when there is no variance in the risk ratings.

In the survey we asked about participant's priorities when installing apps, 48% of the participants prioritized an application's

features over other criteria for choosing apps. The criteria participants could choose from were: ads, permissions, rank, reviews, friends' suggestions, popularity, features, and design.

The Play Store is generally considered easy to use. Would the addition of this information be overwhelming or confusing? We used the NASA TLX instrument to evaluate if there was significant cognitive load in using our experimental marketplace [22]. Only one person reported the workload of this app store as being "quite demanding". Overall 78% of the participants found it easy to work with our Play Store. That is 78% of participants ranked the task of choosing apps as not demanding, not frustrating, low effort, low temporal demand, and easy to accomplish the task. For obvious reasons we did not include the query about the task being physically demanding.

Play Store Rank and App Name	Downloads	Locks
1. Super-Bright LED Flashlight	38	5
3. Color Flashlight	34	5
2. Tiny Flashlight + LED	26	5
4. Brightest Flashlight Free	20	4
10. Flashlight Galaxy S7	16	5
9. Flashlight Galaxy	16	5
5. Brightest LED Flashlight	15	5
11. Flashlight	12	5
6. High-powered Flashlight	11	5
12. Flashlight Widget	7	5
7. FlashLight	6	5
13. Flashlight for HTC	5	5
8. Flashlight	3	5

Table 1: Flashlight Category by order of Downloads in the Experiment: Apps Rank in the Play Store, Downloads in the Experiment, and Safety Rating (locks)

The weighted average risk ratings of the flashlight apps in the Play Store is 4.94. While this number was 4.90 among the choices of participants. The weighted app ratings were 4.52 and 4.51 for the store and the experimental participants, respectively.

Play Store Rank and App Name	Downloads	Locks
1. Google Photos	39	5
8. PhotoDirector Photo Editor App	25	5
5. Photo Lab Picture Editor FX	24	5
9. Gallery	23	5
4. Photo Editor Pro	20	5
11. A+ Gallery Photos & Videos	19	5
5. Photo Collage Editor	17	5
3. PhotoGrid & Photo Collage	15	5
10. Toolwiz Photos - Pro Editor	13	5
6. Photo Editor Collage Maker Pro	9	5
red2. PicsArt Photo Studio & Collage	3	3
7. Phonto - Text on Photos	1	5

Table 2: Photos Category by order of Downloads in the Experiment: Apps Rank in the Play Store, Downloads in the Experiment, and Safety Rating (locks)

The average risk rating of the photos apps in the Play Store is 4.69. The weighted average risk ratings of the choices of participants was 4.97. The weighted app ratings were 4.39 and 4.41. Essentially in the case of photo apps, the distribution of app ratings and risk was such that individuals could mitigate risk without sacrificing any benefits. In Photos, participants chose more secure apps over other more popular apps with more downloads, more familiarity, and more popular design. Specifically, *PicsArt Photo Studio and Collage* were selected by only 3 of our participants in our experiment for the Photos category while it was the second ranked app in terms of number of downloads with this search term in Google Play Store.

Play Store Rank and App Name	Downloads	Locks
2. Fruit Ninja Free	39	5
1. Subway Surfers	23	5
8. Super Smash Jungle World	22	5
5. PAC-MAN	20	5
13. Wheel of Fortune Free Play	16	5
7. Color Switch	15	5
4. Piano Tiles 2™	15	5
3. slither.io	12	5
6. Rolling Sky	11	5
9. Block! Hexa Puzzle	4	5
10. Flip Diving	3	1
16. Battleships - Fleet Battle	2	5
11. Snakes & Ladders King	2	5
13. Board Games	1	5
14. Best Board Games	1	5
12. Checkers	1	5
15. Mancala	1	3

Table 3: Games Category by order of Downloads in the Experiment: Apps Rank in the Play Store, Downloads in the Experiment, and Safety Rating (locks)

The weighted average risk ratings of the games apps in the Play Store was 4.93. The weighted average risk ratings of the choices of participants was also 4.93. The weighted app ratings were 4.43 and 4.34 for the store and the experimental participants, respectively.

Play Store Rank and App Name	Downloads	Locks
1. Weather - The Weather Channel	40	4
2. AccuWeather	31	5
5. Yahoo Weather	27	5
10. MyRadar Weather Radar	27	5
11. Weather Underground	19	5
6. Weather by WeatherBug	16	3
4. Weather& Clock Widget Android	14	4
6. Transparent clock & weather	11	3
12. NOAA Weather Unofficial	7	4
15. Weather Project	5	1
8. Weather, Widget Forecast Radar	3	4
14. Weather Project	2	1
13. iWeather-The Weather Today	2	1
3. Go Weather Forecast & Widgets	5	4
9. Weather	1	4

Table 4: Weather Category by order of Downloads in the Experiment: Apps Rank in the Play Store, Downloads in the Experiment, and Safety Rating (locks)

The weighted average risk ratings of the weather apps in the Play Store was 4.26 and 4.25 for our participants. The weighted average app ratings was 4.39 for both Play Store users and experimental participants.

The trade-off between risk and ratings/downloads is particularly clear in the Weather category, where the most popular app was the same for both cases but those rated as safer were chosen next in our experiment over more popular but more risky apps. The over-privileged apps like *Go Weather Forecast & Widgets* were systematically rejected by our participants.

Comparing results show that in 2 (Photos and Weather) out of 4 categories, participants chose safer apps over apps with more downloads. The Bayesian analysis shows that Games and Photos have a significantly different distribution than those downloaded. Both analyses show that when there was little difference in safety rating or app functionality, our participants‘ results were not distinct from random sample from the Play Store marketplace.

The two categories in which there is a noticeable difference between the results from our experiment and the play store are the categories chosen to answer RQ2 and RQ4. On the other hand, the two categories in which our results were similar to that of the Android marketplace are the two categories related to RQ1 and RQ3 answers. Recall that RQ1 and RQ3 are chosen to check the validity of our experiment as an accurate representation of the android marketplace when the risk ratings do not vary significantly. RQ2 and RQ4 are chosen to verify the variability of the participants' choices from that of the Android marketplace users in the presence of various risk information.

In the mobile market permissions control access to user data and phone functionality. When forced to make a trade-off between risk and benefits, individuals with simplified indicators may choose lower benefit and safer options, as shown in the choices for weather and photos.

It appears however, when there is one choice with far more downloads, that will still be users' top choice.

From the results of this and previous work [34] we can observe that providing risk information did not consistently affect the participants decision about a dominant app. We see that risk ratings did not affect participants' first choices in choosing *Weather - The Weather Channel*. This was still the participants' top choice despite being identified as being more risky.

The average app ratings for the top 5 choices of our participants and the average app ratings for the top 5 choices of the Play Store are quite similar. When app ratings are similar participants choose apps with better safety ratings over apps with more downloads which might mean that with the same quality, participants will choose security over popularity.

6 ANALYSIS

The classic human subjects experiment is an A/B test which gives two sets of data. The means of the two groups are compared, usually using a post-hoc Tukey (pairwise comparison). We include these results for each category for the ease of comparison with other work. The Kruskal-Wallis shows the significance of Differences in Weighted Means of Risk Ratings for the four categories which are, in order, 0.005 for Games; 0.53 for Flashlights; 0.02 for Photos; and 0.28 for Weather. The results of this comparison shows the significance of the differences between the mean risk rating of apps chosen by those using our experimental Play Store and the mean risk rating of apps chosen through the actual Android Play Store.

These results are not a substantive analysis, but like a bar chart, provide an illustrative introduction to the analysis below. Not surprisingly, the Flashlight category shows the smallest difference as the most popular Flashlights already have safety ratings of 5. Thus the difference between those with and without preferences information was not significantly different. A naive means comparison would indicate that the risk rating actually decreased. However, the analyses below show that in fact our participants distribution of flashlight choices was statistically indistinguishable from those in the larger Play Store.

This result and the following analysis (along with participant recruitment from public places while excluding computer scientists) provides evidence that our sample was not somehow exogenously distinct from the Android marketplace.

As was shown in the data description above, Weather was notable in that a few low security apps which are popular in the Play Store were rejected in the experiment. We also noted that The Weather Channel remained highly rated despite its relatively aggressive use of permissions and associated lower rating.

The Photos group of applications was significantly different. The Photo apps show difference in terms of functionality. With Flashlights and Weather reports, there is limited difference in the apps, however in Photos category, functionality of various Photo management apps can be different.

Figures 4- 7 show the results from a Bayesian analysis. The value of a Bayesian analysis in this case is that there is rich information about the distributions of the apps individuals actually choose in the Android Play Store, as opposed to differences in means. Our goal is to compare the behavior of the participants using this augmented interaction with the behavior of participants who use a normal Play Store. The use of the overall market as a comparison

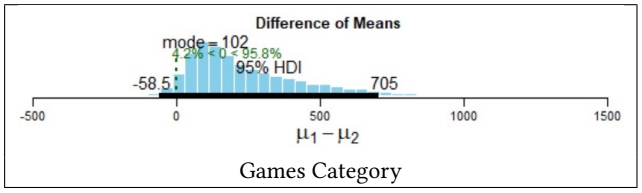

Figure 4: Regions of Practical Equivalence - Games (RQ3)

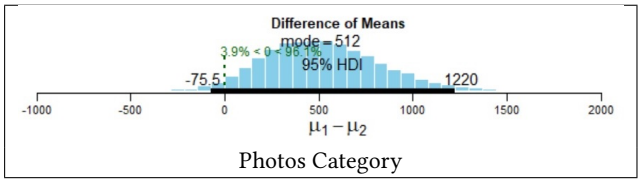

Figure 5: Regions of Practical Equivalence - Photos (RQ4)

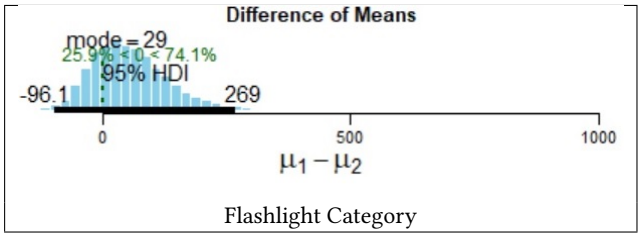

Figure 6: Regions of Practical Equivalence - Flashlight(RQ1)

is grounded in the objective of developing a system that alters the order and selection relative to the normal Android marketplace. From a Bayesian perspective, we have high confidence and a well-informed prior distribution of the selection of apps under the normal condition. We use that distribution of known app selections as our prior, we then test to determine if the apps selected in our experiment are likely to have arisen given that known distribution.

Consider that we are seeking a bias in participants' decision-making. Imagine if we were testing dice to see if these were biased. We would take the tremendous prior information about the patterns of dice (e.g., one in six to get a particular number on a six-sided die). We would then roll our die and compare this against the prior distribution to determine if there were biases. Here, we are seeking to determine if there is bias when risk varies in a category.

We calculated a region of practical equivalence (ROPE) based on a Highest Density Interval (HDI) of 95%. The comparison is between the behavior of our experimental sample as opposed to the behavior of people using the normal Android Play Store. (Other terms used for ROPE include indifference zone, smallest effect size of interest, or clinical equivalence.) The results are similar to those of the means comparison above. There are no assumptions about normality, or distribution, but unlike the means test a Bayesian approach evaluates all the distributional parameters (e.g., standard deviation) not only means.

In the case of the Flashlight apps, there is almost no variance of functionality. There was very little variance in privacy ratings. The

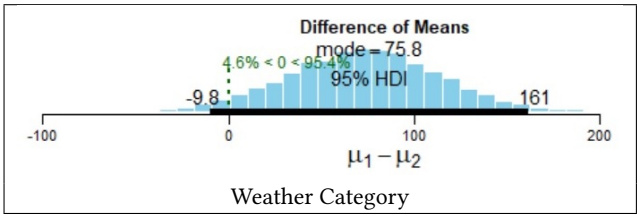

Figure 7: Regions of Practical Equivalence - Weather (RQ2)

choices of Flashlight app in our experimentation group was statistically indistinguishable from a random sample of the selections made in the larger Play Store.

Specifically, the flashlight region of equivalence is such that perfect random chance, 50%, is nearly at the center of this distribution. The Flashlight selections are indistinguishable from the true Play Store selections which verifies that in the absence of differing risk ratings our participants' choices were indistinguishable from those of a random sample of Android users.

The Weather result is significant. Unlike with a single variable, it is possible to observe the distribution of likely parameters. Thus 95.4% > 95%, the graph illustrates the distribution over the parameter values. So by observing values within the HDI there is some difference between the distribution of apps selected by the experimental group and those normally selected in the Android Play Store.

The dominance of the most popular Weather app, with a risk rating of four, results in a slight skewing of the results. The overall mean of the difference between Weather apps with risk showin in Figure 7 shows very little overlap between the distribution of selected weather apps under a Bayesian comparison with the Play Store and by our participants. Kruskal-Wallis difference in means had a p value of 0.28. (Note that Kruskal-Wallis examined the difference of the means, while a Bayesian approach considers the likelihood in the context of all the possible distributions.) That is, the means difference was not significant however the overall distribution can be statistically distinguished from random chance.

The difference between Photos and Weather shows that both are significant, but that difference has decreased with this analysis compared with the means comparison above. Instead of 4.8% of the possible parameters which could explain the results being outside of the ROPE, the HDI is 96.1%. Thus we can be confident that in the case of Photos, there was an effect from the addition of the lock icons to distinguish between apps with higher and lower risk.

The Games category result also shows less significance than Photos and Weather category. Notice that the values of credible parameters are highly skewed, both in the prior and in our results. This reflects the uncertainty discussed above, where we note that there is a wider range of Games with the same safety ratings.

7 DISCUSSION

In general, it is necessary for consumers to have clear and correct information at the time of purchase for a market to function [4]. We applied these long-established principles to the design of decision support in the selection of apps.

The very existence of a demand for privacy is contested. If there is no desire for privacy, then increasing the transparency of a decision will not change the decision. If the manner in which privacy information is provided is not usable, then it will similarly be ignored. We built on previous work in usable security and decision-making in security to create a theoretically grounded interaction. The research questions illustrate that there was no a priori certainty that additional information about privacy would change download decisions.

Previous research has examples where information about risk has been unwelcome or counter-productive. Studies in risk communication have shown that individuals find risk more acceptable if the exposure to the risk is voluntary; and the individual exposed is capable of mitigating the impact of it [19]. That is, shifting the nexus of control may increase aggregate risk-taking. This response is called the 'control dilemma' [10]; that is the perception of control increases data sharing and longitudinal exposure may increase mitigating activity [39]. The fact that individuals have already chosen an Android device may imply that the endowment effect impinges their decisions. This would mitigate risk concerns in purchases, as shown in the case of known privacy breaches [3].

In order to address these questions in terms of mobile apps, we asked the research questions described above. First (RQ1) can we confirm that our group of participants make choices that are indistinguishable from the Android market users as a whole when presented with apps with the same kind of functionality and the same risk ratings? We used the Flashlight category for this purpose. In this category the functionality of all apps is arguably the same and the safety ratings are all identical.

Our next question (RQ2) is if the participants would make different choices compared to that of Android market users in the presence of various risk ratings with the same app functionality. For this question, we used weather category. In this category, functionalities are similar but the risk ratings varied significantly.

Next, we ask (in RQ3) if our participants' choice is indistinguishable from Android market users when the functionality is different but the risk ratings vary slightly. We use games as the category to answer this question.

Our last question (RQ4) is that how comparable are our participants' choices to that of Android market users when we have relatively varying functionalities and quite different risk ratings. We used the photos category in which functionality was less similar than that of the apps in the weather category. The risk ratings also vary.

If there is no change in ranking between these four categories in our experiment versus the Play Store, this will support the contention that individuals want free apps, and app benefits outweigh app risks.

If there were a greater change when the functionality is the same, this indicates that people would choose privacy if there is no loss in functionality.

If the largest difference in the distributions occurs when there is the option to trade functionality for privacy, this indicates that people would choose safer apps even if there is a corresponding loss of functionality.

8 CONCLUSIONS

In this paper we developed a Play Store interaction to test the possibility that risk communication affects app choices. We did this by providing information that allowed consumers to easily distinguish between an app that was high risk (indicated by low safety, and few locks) or low risk (indicated by more safety, and more locks). Our contributions are an indication that if consumers are provided information about risk, and marketplace options that allow them to choose risk over popularity, enough people will do so as to change the marketplace dynamics.

Specifically, we altered the Play Store to include risk ratings. We then recruited sixty participants to select four apps in each of four categories. We compared the resulting selections with the ratings to the prior distribution provided by the real world. Specifically, we compared the risk ratings and the user ratings of the apps selected by our population using the distribution of selections as an informed prior. In terms of the Flashlight app, our risk ratings had a very low standard deviation and a high mean. There was not an effective option for a risk/benefit trade-off. The statistical analysis indicated that our Flashlight results were indistinguishable from a random sample from the larger Android population.

In Weather apps the ranking and selections were different. However, the overall mean was not statistically different. The Weather Channel, with a lower safety rating, remained by far the most popular app. The distribution of selections was not what was predicted by the prior. So the mean did not change, but the ranking in marketplace show evidence of change.

The Photo result showed a difference in means and in rankings, with participants choosing higher safety as opposed to more popularity.

The Games results also show less significance than Photos and Weather category which can be a result of relatively similar risk ratings.

Future work in this arena includes the addition of other types of interactions, to verify the changes. In addition, future work will include a comparison of the our tablet-based Play Store and the same interaction of Play Store used on MTurk. Our goal is to embed the Play Store into a large number of participants' phones and observe changes. We seek a partner with this infrastructure. We were motivated by a belief that our risk communication could result in safer app choices, overall decrease in information exfiltration and over permissioning. This would decrease the aggregate vulnerability of the Android ecosystem.

ACKNOWLEDGEMENTS

We would like to acknowledge the assistance of Prashanth Rajivan who provided valuable feedback, as well as Yeesoo Chae and Rachel Huss, who assisted in implementation of the experiment. This research was supported in part by the National Science Foundation under CNS 1565375, Cisco Research Support #591000, and the Comcast Innovation Fund. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the the US Government, the National Science Foundation, Cisco, Comcast, or Indiana University.

REFERENCES

[1] M. S. Ackerman and L. Cranor. Privacy critics: Ui components to safeguard users' privacy. In *CHI '99 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '99, pages 258–259, New York, NY, USA, 1999. ACM.

[2] A. Acquisti, L. Brandimarte, and G. Loewenstein. Privacy and human behavior in the age of information. *Science*, 347(6221):509–514, 2015.

[3] S. Afroz, A. C. Islam, J. Santell, A. Chapin, and R. Greenstadt. How privacy flaws affect consumer perception. In *Socio-Technical Aspects in Security and Trust (STAST), 2013 Third Workshop on*, pages 10–17. IEEE, 2013.

[4] G. Akerlof. The market for "lemons": Quality uncertainty and the market mechanism. In *Essential Readings in Economics*, pages 175–188. Springer, 1995.

[5] C. Amrutkar, K. Singh, A. Verma, and P. Traynor. Vulnerableme: Measuring systemic weaknesses in mobile browser security. *ICISS*, 7671:16–34, 2012.

[6] S. Arzt, S. Rasthofer, C. Fritz, E. Bodden, A. Bartel, J. Klein, Y. Le Traon, D. Octeau, and P. McDaniel. Flowdroid: Precise context, flow, field, object-sensitive and lifecycle-aware taint analysis for android apps. *Acm Sigplan Notices*, 49(6):259–269, 2014.

[7] D. Barrera, H. G. Kayacik, P. C. van Oorschot, and A. Somayaji. A methodology for empirical analysis of permission-based security models and its application to android. In *Proceedings of the 17th ACM conference on Computer and communications security*, pages 73–84. ACM, 2010.

[8] K. Benton, L. J. Camp, and V. Garg. Studying the effectiveness of android application permissions requests. In *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2013 IEEE International Conference on*, pages 291–296. IEEE, 2013.

[9] R. Böhme and J. Grossklags. The security cost of cheap user interaction. In *Proceedings of the 2011 workshop on New security paradigms workshop*, pages 67–82. ACM, 2011.

[10] L. Brandimarte, A. Acquisti, and G. Loewenstein. Misplaced confidences privacy and the control paradox. *Social Psychological and Personality Science*, 4(3):340–347, 2013.

[11] I. Burguera, U. Zurutuza, and S. Nadjm-Tehrani. Crowdroid: behavior-based malware detection system for android. In *Proceedings of the 1st ACM workshop on Security and privacy in smartphones and mobile devices*, pages 15–26. ACM, 2011.

[12] W. S. Bush and S. P. Vecera. Differential effect of one versus two hands on visual processing. *Cognition*, 133(1):232–237, 2014.

[13] S. Egelman, L. F. Cranor, and J. Hong. You've been warned: an empirical study of the effectiveness of web browser phishing warnings. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1065–1074. ACM, 2008.

[14] W. Enck, P. Gilbert, S. Han, V. Tendulkar, B.-G. Chun, L. P. Cox, J. Jung, P. McDaniel, and A. N. Sheth. Taintdroid: an information-flow tracking system for realtime privacy monitoring on smartphones. *ACM Transactions on Computer Systems (TOCS)*, 32(2):5, 2014.

[15] W. Enck, D. Octeau, P. McDaniel, and S. Chaudhuri. A study of android application security. In *USENIX security symposium*, volume 2, page 2, 2011.

[16] A. P. Felt, E. Chin, S. Hanna, D. Song, and D. Wagner. Android permissions demystified. In *Proceedings of the 18th ACM conference on Computer and communications security*, pages 627–638. ACM, 2011.

[17] A. P. Felt, E. Ha, S. Egelman, A. Haney, E. Chin, and D. Wagner. Android permissions: User attention, comprehension, and behavior. In *Proceedings of the Eighth Symposium on Usable Privacy and Security*, page 3. ACM, 2012.

[18] A. Forget, S. Pearman, J. Thomas, A. Acquisti, N. Christin, L. F. Cranor, S. Egelman, M. Harbach, and R. Telang. Do or do not, there is no try: user engagement may not improve security outcomes. In *Symposium on Usable Privacy and Security (SOUPS)*, 2016.

[19] V. Garg and J. Camp. End user perception of online risk under uncertainty. In *System Science (HICSS), 2012 45th Hawaii International Conference on*, pages 3278–3287. IEEE, 2012.

[20] J. Grossklags and A. Acquisti. When 25 cents is too much: An experiment on willingness-to-sell and willingness-to-protect personal information. In *WEIS*, 2007.

[21] M. A. Harris, R. Brookshire, K. Patten, and B. Regan. Mobile application installation influences: have mobile device users become desensitized to excessive permission requests? In *Proceedings of the Twentieth Americas Conference on Information Systems (AMCIS 2015)*, pages 13–15, 2015.

[22] S. G. Hart and L. E. Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. *Advances in psychology*, 52:139–183, 1988.

[23] S. G. Hart and L. E. Stavenland. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock and N. Meshkati, editors, *Human Mental Workload*, chapter 7, pages 139–183. Elsevier, 1988.

[24] P. G. Kelley, L. F. Cranor, and N. Sadeh. Privacy as part of the app decision-making process. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3393–3402. ACM, 2013.

[25] S. Kolimi, F. Zhu, and S. Carpenter. Contexts and sharing/not sharing private information. In *Proceedings of the 50th Annual Southeast Regional Conference*, pages 292–297. ACM, 2012.

[26] M. Korayem, R. Templeman, D. Chen, D. Crandall, and A. Kapadia. Screenavoider: Protecting computer screens from ubiquitous cameras. *arXiv preprint arXiv:1412.0008*, 2014.

[27] L. Kraus, I. Wechsung, and S. Möller. Using statistical information to communicate android permission risks to users. In *Socio-Technical Aspects in Security and Trust (STAST), 2014 Workshop on*, pages 48–55. IEEE, 2014.

[28] I. Liccardi, J. Pato, D. J. Weitzner, H. Abelson, and D. De Roure. No technical understanding required: Helping users make informed choices about access to their personal data. In *Proceedings of the 11th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*, pages 140–150. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2014.

[29] J. Lin, S. Amini, J. I. Hong, N. Sadeh, J. Lindqvist, and J. Zhang. Expectation and purpose: understanding users' mental models of mobile app privacy through crowdsourcing. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pages 501–510. ACM, 2012.

[30] mar-v in. Blankstore, Jul 2017. https://github.com/mar-v-in/BlankStore.

[31] G. R. Milne and M. J. Culnan. Strategies for reducing online privacy risks: Why consumers read (or don't read) online privacy notices. *Journal of Interactive Marketing*, 18(3):15–29, 2004.

[32] A. Morton. "all my mates have got it, so it must be okay": Constructing a richer understanding of privacy concerns—an exploratory focus group study. In *Reloading Data Protection*, pages 259–298. Springer, 2014.

[33] E. Pan, J. Ren, M. Lindorfer, C. Wilson, and D. Choffnes. Panoptispy: Characterizing audio and video exfiltration from android applications. *Proceedings on Privacy Enhancing Technologies*, 2018(4):33–50, 2018.

[34] P. Rajivan and J. Camp. Influence of privacy attitude and privacy cue framing on android app choices. In *Authentication Workshop of the 12th Symposium on Usable Privacy and Security*. USENIX Association, 2016.

[35] J. Sadeh and J. I. Hong. Modeling users' mobile app privacy preferences: Restoring usability in a sea of permission settings. In *Symposium on Usable Privacy and Security (SOUPS)*, volume 40, 2014.

[36] R. Schlegel, A. Kapadia, and A. J. Lee. Eyeing your exposure: quantifying and controlling information sharing for improved privacy. In *Proceedings of the Seventh Symposium on Usable Privacy and Security*, page 14. ACM, 2011.

[37] A.-D. Schmidt, R. Bye, H.-G. Schmidt, J. Clausen, O. Kiraz, K. A. Yuksel, S. A. Camtepe, and S. Albayrak. Static analysis of executables for collaborative malware detection on android. In *Communications, 2009. ICC'09. IEEE International Conference on*, pages 1–5. IEEE, 2009.

[38] W. Shin, S. Kiyomoto, K. Fukushima, and T. Tanaka. Towards formal analysis of the permission-based security model for android. In *Wireless and Mobile Communications, 2009. ICWMC'09. Fifth International Conference on*, pages 87–92. IEEE, 2009.

[39] F. Stutzman, R. Gross, and A. Acquisti. Silent listeners: The evolution of privacy and disclosure on facebook. *Journal of privacy and confidentiality*, 4(2):2, 2013.

[40] J. Y. Tsai, S. Egelman, L. Cranor, and A. Acquisti. The effect of online privacy information on purchasing behavior: An experimental study. *Information Systems Research*, 22(2):254–268, 2011.

[41] T. Vila, R. Greenstadt, and D. Molnar. Why we can't be bothered to read privacy policies models of privacy economics as a lemons market. In *Proceedings of the 5th international conference on Electronic commerce*, pages 403–407. ACM, 2003.

[42] H. Wang, J. Hong, and Y. Guo. Using text mining to infer the purpose of permission use in mobile apps. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 1107–1118. ACM, 2015.

[43] X. Wang, F. Du, X. He, and K. Zhang. Enhanced spatial stimulus–response mapping near the hands: The simon effect is modulated by hand-stimulus proximity. *Journal of Experimental Psychology*, 40(6):2252, 2014.

[44] L. Yang, N. Boushehrinejadmoradi, P. Roy, V. Ganapathy, and L. Iftode. Short paper: enhancing users' comprehension of android permissions. In *Proceedings of the second ACM workshop on Security and privacy in smartphones and mobile devices*, pages 21–26. ACM, 2012.