

# The Feasibility of Dynamically Granted Permissions: Aligning Mobile Privacy with User Preferences

Primal Wijesekera<sup>1</sup>, Arjun Baokar<sup>2</sup>, Lynn Tsai<sup>2</sup>, Joel Reardon<sup>2</sup>,  
Serge Egelman<sup>2</sup>, David Wagner<sup>2</sup>, and Konstantin Beznosov<sup>1</sup>

<sup>1</sup>University of British Columbia, Vancouver, Canada,  
{primal,beznosov}@ece.ubc.ca

<sup>2</sup>University of California, Berkeley, Berkeley, USA,  
{arjunbaokar,lynntsai,joel.reardon}@berkeley.edu, {egelman,daw}@cs.berkeley.edu

**Abstract**—Current smartphone operating systems regulate application permissions by prompting users on an ask-on-first-use basis. Prior research has shown that this method is ineffective because it fails to account for context: the circumstances under which an application first requests access to data may be vastly different than the circumstances under which it subsequently requests access. We performed a longitudinal 131-person field study to analyze the contextuality behind user privacy decisions to regulate access to sensitive resources. We built a classifier to make privacy decisions on the user’s behalf by detecting when context has changed and, when necessary, inferring privacy preferences based on the user’s past decisions and behavior. Our goal is to automatically grant appropriate resource requests without further user intervention, deny inappropriate requests, and only prompt the user when the system is uncertain of the user’s preferences. We show that our approach can accurately predict users’ privacy decisions 96.8% of the time, which is a four-fold reduction in error rate compared to current systems.

## I. INTRODUCTION

One of the roles of a mobile application platform is to help users avoid unexpected or unwanted use of their personal data [12]. Mobile platforms currently use permission systems to regulate access to sensitive resources, relying on user prompts to determine whether a third-party application should be granted or denied access to data and resources. One critical caveat in this approach, however, is that mobile platforms seek the consent of the user the first time a given application attempts to access a certain data type and then enforce the user’s decision for all subsequent cases, regardless of the circumstances surrounding each access. For example, a user may grant an application access to location data because she is using location-based features, but by doing this, the application can subsequently access location data for behavioral advertising, which may violate the user’s preferences.

Earlier versions of Android (5.1 and below) asked users to make privacy decisions during application installation as an all-or-nothing ultimatum (ask-on-install): either all requested permissions are approved or the application is not installed. Previous research showed that few people read the requested permissions at install-time and even fewer correctly understood them [17]. Furthermore, install-time permissions do not present users with the context in which those permission will

be exercised, which may cause users to make suboptimal decisions not aligned with their actual preferences. For example, Egelman et al. observed that when an application requests access to location data without providing context, users are just as likely to see this as a signal for desirable location-based features as they are an invasion of privacy [11]. Asking users to make permission decisions at runtime—at the moment when the permission will actually be used by the application—provides more context (i.e., what they were doing at the time that data was requested) [15]. However, due to the high frequency of permission requests, it is not feasible to prompt the user every time data is accessed [43].

In iOS and Android M, the user is now prompted at runtime the first time an application attempts to access one of a set of “dangerous” permission types (e.g., location, contacts, etc.). This *ask-on-first-use* (AOFU) model is an improvement over ask-on-install (AOI). Prompting users the first time an application uses one of the designated permissions gives users a better sense of context: their knowledge of what they were doing when the application first tried to access the data should help them determine whether the request is appropriate. Despite that, Wijesekera et al. showed that AOFU fails to meet user expectations over half the time. This is because AOFU does not account for the varying contexts of future requests [43].

The notion of *contextual integrity* suggests that many permission models fail to protect user privacy because they fail to account for the context surrounding data flows [34]. That is, privacy violations occur when sensitive resources are used in ways that defy users’ expectations. We posit that more effective permission models must focus on whether resource accesses are likely to defy users’ expectations in a given context—not simply whether the application was authorized to receive data the first time it asked for it. Thus, the challenge for system designers is to correctly infer when the context surrounding a data request has changed, and whether the new context is likely to be deemed “appropriate” or “inappropriate” for the given user. Dynamically regulating data access based on the context requires more user involvement to understand users’ contextual preferences. If users are asked to make privacy decisions too frequently, or under circumstances that are seen as low-risk, they may become habituated to future,

more serious, privacy decisions. On the other hand, if users are asked to make too few privacy decisions, they may find that the system has acted against their wishes. Thus, our goal is to automatically determine *when* and under *what* circumstances the system presents users with runtime prompts.

To this end, we collected real-world Android usage data in order to explore whether we could infer users' future privacy decisions based on their past privacy decisions, contextual circumstances surrounding applications' data requests, and users' behavioral traits. We conducted a field study where 131 participants used Android phones that were instrumented to gather data over an average of 32 days per participant. Also, their phones periodically prompted them to make privacy decisions when applications used sensitive permissions, and we logged their decisions. Overall, participants wanted to block 60% of these requests. We found that AOFU yields 84% accuracy, i.e., its policy agrees with participants' prompted responses 84% of the time. AOI achieves only 25% accuracy.

We designed new techniques that use machine learning to automatically predict how users would respond to prompts, so that we can avoid prompting them in most cases, thereby reducing user burden. Our classifier uses the user's past decisions in similar situations to predict their response to a particular permission request. The classifier outputs a prediction and a confidence score; if the classifier is sufficiently confident, we use its prediction, otherwise we prompt the user for their decision. We also incorporate information about the user's behavior in other security and privacy situations to make inferences about their preferences: whether they have a screen lock activated, how often they visit HTTPS websites, and so on. We show that our scheme achieves 96.8% accuracy (a 4× reduction in error rate over AOFU) with significantly less user involvement than the *status quo*.

The specific contributions of our work are the following:

- We conducted the first known large-scale study on quantifying the effectiveness of ask-on-first-use permissions.
- We show that a significant portion of the studied participants make contextual decisions on permissions—the foreground application and the visibility of the permission-requesting application are strong cues participants used to make contextual decisions.
- We show how a machine-learned model can incorporate context and better predict users' privacy decisions.
- To our knowledge, we are the first to use passively observed traits to infer future privacy decisions on a case-by-case basis at runtime.

## II. RELATED WORK

There is a large body of work demonstrating that install-time prompts fail because users do not understand or pay attention to them [19], [23], [42]. When using install-time prompts, users often do not understand which permission types correspond to which sensitive resources and are surprised by the ability of background applications to collect information [17], [22], [41]. Applications also transmit a large amount of location or other sensitive data to third parties without

user consent [12]. When possible risks associated with these requests are revealed to users, their concerns range from being annoyed to wanting to seek retribution [16].

To mitigate some of these problems, systems have been developed to track information flows across the Android system [12], [18], [24] or introduce finer-grained permission control into Android [2], [21], [39], but many of these solutions increase user involvement significantly, which can lead to habituation. Additionally, many of these proposals are useful only to the most-motivated or technically savvy users. For example, many such systems require users to configure complicated control panels, which many are unlikely to do [45]. Other approaches involve static analysis in order to better understand how applications *could* request information [4], [8], [14], but these say little about how applications *actually* use information. Dynamic analysis improves upon this by allowing users to see how often this information is requested in real time [12], [40], [43], but substantial work is likely needed to present that information to average users in a meaningful way. Solutions that require user interruptions need to also minimize user intervention in order to prevent habituation.

Other researchers have developed recommendation systems to recommend applications based on users' privacy preferences [46], or detect privacy violations and suggest preferences based on crowdsourcing [1], [27], but such approaches often do not take individual user differences into account without significant user intervention. Systems have also been developed to predict what users would share on mobile social networks [7], which suggests that future systems could potentially infer what information users would be willing to share with third-party applications. By requiring users to self-report privacy preferences, clustering algorithms have been used to define user privacy profiles even in the face of diverse preferences [26], [38]. However, researchers have found that the order in which information is requested has an impact on prediction accuracy [44], which could mean that such systems are only likely to be accurate when they examine actual user behavior over time (as opposed to one-time self-reports).

Liu et al. clustered users by privacy preferences and used ML techniques to predict whether to allow or deny an application's request for sensitive user data [29]. Their dataset, however, was collected from a set of highly privacy-conscious individuals: those who choose to install a permission-control mechanism. Furthermore, the researchers removed "conflicting" user decisions, in which a user chose to deny a permission for an application, and then later chose to allow it. These conflicting decisions, however, do not represent noisy data. They occur nearly 50% of the time in the real world [43], and accurately reflect the nuances of user privacy preferences. Models must therefore account for them. In fact, previous work found that users commonly reassess privacy preferences after usage [3]. Liu et al. also expect users to make 10% of permission decisions manually, which, based on field study results from Wijesekera et al., would result in being prompted every three minutes [43]. This is obviously impractical. Our goal is to design a system that can automatically make decisions on

behalf of users, that accurately models their preferences, while also not over-burdening them with repeated requests.

Closely related to this work, Liu et al. [28] performed a field study to measure the effectiveness of a Privacy Assistant that offers recommendations to users on privacy settings that they could adopt based on each user’s privacy profile—the privacy assistant predicts what the user might want based on the inferred privacy profile and static analysis of the third-party application. While this approach increased user awareness on resource usage, the recommendations are static: they do not consider each application’s access to sensitive data on a case-by-case basis. Such a coarse-grained approach goes against previous work suggesting that people do want to vary their decisions based on contextual circumstances [43]. A blanket approval or denial of a permission to a given application carries a considerable risk of privacy violations or loss of desired functionality. In contrast, our work uses dynamic analysis to infer the appropriateness of each given request by considering the surrounding contextual cues and how the user has behaved in similar situations in the past. As with Liu et al., their dataset was also collected from privacy-conscious and considerably tech-savvy individuals, which may limit the generalization of their results. The field study we conduct in our work uses a more representative sample.

Nissenbaum’s theory of contextual integrity suggests that permission models should focus on information flows that are likely to defy user expectations [34]. There are three main components involved in deciding the appropriateness of a flow [6]: the context in which the resource request is made, the role played by the requesting application under the current context, and the type of resource being accessed. Neither previous nor currently deployed permission models take all three factors into account. This model could be used to improve permission models by automatically granting access to data when the system determines that it is appropriate, denying access when it is inappropriate, and prompting the user only when a decision cannot be made automatically, thereby reducing user burden.

*Access Control Gadgets* (ACGs) were proposed as a mechanism to tie sensitive resource access to certain UI elements [32], [35]–[37]. Authors posit that such an approach will increase user expectations, as a significant portion of participants expected a UI interaction before a sensitive resource usage, giving users an implicit mechanism to control access and increasing awareness on resource usage. The biggest caveat in this approach is that tying a UI interaction to each sensitive resource access is impossible in practice because resources are accessed at a high frequency [43], and because many legitimate resource accesses occur without user initiation [15].

Wijesekera et al. performed a field study [43] to operationalize the notion of “context,” to allow an operating system to differentiate between appropriate and inappropriate data requests by a single application for a single data type. They found that users’ decisions to allow a permission request significantly correlated with that application’s visibility. They posit that this visibility is a strong contextual cue that influences users’

Permission Type	Activity
ACCESS_WIFI_STATE	View nearby SSIDs
NFC	Communicate via NFC
READ_HISTORY_BOOKMARKS	Read users’ browser history
ACCESS_FINE_LOCATION	Read GPS location
ACCESS_COARSE_LOCATION	Read network-inferred location (i.e., cell tower and/or WiFi)
LOCATION_HARDWARE	Directly access GPS data
READ_CALL_LOG	Read call history
ADD_VOICEMAIL	Read call history
READ_SMS	Read sent/received/draft SMS
SEND_SMS	Send SMS
*INTERNET	Access Internet when roaming
*WRITE_SYNC_SETTINGS	Change application sync settings when roaming

TABLE I  
FELT ET AL. PROPOSED GRANTING A SELECT SET OF 12 PERMISSIONS AT RUNTIME SO THAT USERS HAVE CONTEXTUAL INFORMATION TO INFER WHY THE DATA MIGHT BE NEEDED [15]. OUR INSTRUMENTATION OMITTS THE LAST TWO PERMISSION TYPES (INTERNET & WRITE\_SYNC\_SETTINGS) AND RECORDS INFORMATION ABOUT THE OTHER 10.

responses to permission prompts. They also observed that privacy decisions were highly nuanced, demonstrating that a one-size-fits-all model is unlikely to be sufficient; a given information flow may be deemed appropriate by one user but not by another user. They recommended applying machine learning in order to infer individual users’ privacy preferences.

To achieve this, research is needed to determine what factors affect user privacy decisions and how to use those factors to make privacy decisions on the user’s behalf. While we cannot automatically capture everything involved in Nissenbaum’s notion of *context*, we can try to detect when context has likely changed (insofar as to decide whether a different privacy decision should be made for the same application and data type), by seeing whether the circumstances surrounding a data request are similar to previous requests.

### III. METHODOLOGY

We collected data from 131 participants to understand what factors could be used to infer whether a permission request is likely to be deemed appropriate by the user.

Previous work by Felt et al. made the argument that certain permissions are appropriate for runtime prompts, because they protect sensitive resources and because viewing the prompt at runtime imparts additional contextual information about why an application might need the permission [15]. Similarly, Thompson et al. showed that other permission requests could be replaced with audit mechanisms, because they represent either reversible changes or are sufficiently low risk to not warrant habituating the user to prompts [41]. We collected information about 10 of the 12 permissions Felt et al. suggest are best-suited for runtime prompts. We omitted INTERNET and WRITE\_SYNC\_SETTINGS, because those permissions only warrant runtime prompts if the user is roaming and we did not expect any participant to be roaming during the study period, and focused on the remaining 10 permission types (Table I). While there are many other sensitive permissions beyond this

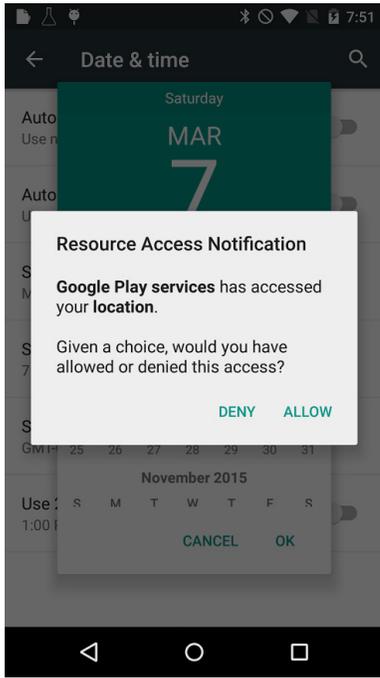


Fig. 1. A screenshot of an ESM prompt.

set, Felt et al. concluded that the others are best handled by other mechanisms (e.g., install-time prompts, ACGs, etc.).

We used the Experience Sampling Method (ESM) to collect ground truth data about users’ privacy preferences [20]. ESM involves repeatedly questioning participants *in situ* about a recently observed event; in this case, we probabilistically asked them about an application’s recent access to data on their phone, and whether they would have permitted it if given the choice. We treated participants’ responses to these ESM probes as our main dependent variable (Figure 1).

We also instrumented participants’ smartphones to obtain data about their privacy-related behaviors and the frequency with which applications accessed protected resources. The instrumentation required a set of modifications to the Android operating system and flashing a custom Android version onto participants’ devices. To facilitate such experiments, the University of Buffalo offers non-affiliated academic researchers access to the PhoneLab panel [33], which consists of more than 200 participants. All of these participants had LG Nexus 5 phones running Android 5.1.1 and the phones were periodically updated over-the-air (OTA) with custom modifications to the Android operating system. Participants can decide when to install the OTA update, which marks their entry into new experiments. During our experiment period, different participants installed the OTA update with our instrumentation at different times, thus we have neither data on all PhoneLab participants nor data for the entire period. Our OTA update was available to participants for a period of six weeks, between February 2016 and March 2016. At the end of the study period, we emailed participants a link to an exit survey to collect demographic

Type	Event Recorded
Behavioral Instrumentation	Changing developer options
	Opening/Closing security settings
	Changing security settings
	Enabling/Disabling NFC
	Changing location mode
	Opening/Closing location settings
	Changing screen-lock type
	Use of two factor authentication
	Log initial settings information
	User locks the screen
	Screen times out
	App locks the screen
	Audio mode changed
	Enabling/Disabling speakerphone
	Connecting/Disconnecting headphones
	Muting the phone
	Taking an audio call
	Taking a picture (front- vs. rear-facing)
	Visiting an HTTPS link in Chrome
	Responding to a notification
Unlocking the phone	
Runtime Information	An application changing the visibility
	Platform switches to a new activity
Permission Requests	An app requests a sensitive permission
	ESM prompt for a selected permission

TABLE II  
INSTRUMENTED EVENTS THAT FORM OUR FEATURE SET

information. Our study received institutional review board (IRB) approval.<sup>1</sup>

#### A. Instrumentation

The goal of our instrumentation was to collect as much runtime and behavioral data as could be observed from the Android platform, with minimal performance cost. We collected three categories of data: behavioral information, runtime information, and user decisions. We made no modifications to any third-party application code; our dynamic analysis techniques could be used on any third-party Android application.

Table II contains the complete list of behavioral and runtime events our instrumentation recorded. The behavioral data fell under several categories, all chosen based on several hypotheses that we had about the types of behaviors that might correlate with privacy preferences: web-browsing habits, screen locking behavior, third-party application usage behavior, audio preferences, call habits, camera usage patterns, and behavior related to security settings. For example, we hypothesized that someone who manually locks their device screen are more privacy-conscious than someone who lets it time out.

We also collected runtime information about the context of each permission request, including the visibility of the requesting application at the time of request, what the user was doing when the request was made (i.e., the name of the foreground application), and the exact Android API function invoked by the application to determine what information was requested.. The visibility of an application reflects the extent to which the

<sup>1</sup>Approved by the UC Berkeley IRB under protocol #2013-02-4992

user was likely aware that the application was running; if the application was in the foreground, the user had cues that the application was running, but if it was in the background, then the user was likely not aware that the application was running and therefore might find the permission request unexpected—some background services can still be visible to the user due to on-screen notification or other cues that could be perceptible. We monitored processes’ memory priority levels to determine the visibility of all Android processes. We also collected information about which Android `Activity` was active in the application.<sup>2</sup>

Once per day we *probabilistically* selected one of these permission requests and prompted the user about them at runtime (Figure 1). We used weighted reservoir sampling to select a permission request to prompt about. We weight the combination of *application*, *permission*, *visibility* based on their frequency of occurrence seen by the instrumentation; the most-frequent combination has a higher probability of being shown to participants using ESM. We prompted participants a maximum of three times for each unique combination. We tuned the wording of the prompt to make it clear that the request had just occurred and their response would not affect the system (a deny response would not actually deny data). These responses serve as the ground truth for all the analysis mentioned in the remainder of the paper.

The intuition behind using weighted reservoir sampling is to focus more on the frequently occurring permission requests over rare ones. Common permission requests contribute most to user habituation due to their high frequency. Thus, it is more important to learn about user privacy decisions on highly frequent permission requests over the rare ones, which might not risk user habituation or annoyance (and the context of rare requests may be less likely to change).

### B. Exit Survey

At the end of our data collection period, PhoneLab staff emailed participants a link to our online exit survey, which they were incentivized to complete with a raffle for two \$100 Amazon gift cards. The survey gathered demographic information and qualitative information on their privacy preferences. Of the 203 participants in our experiment, 53 fully completed the survey, and another 14 partially completed it. Of the 53 participants to fully complete the survey, 21 were male, 31 were female, and 1 undisclosed. Participants ranged from 20 to 72 years of age ( $\mu = 40.83$ ,  $\sigma = 14.32$ ). Participants identified themselves as 39.3% staff, 32.1% students, 19.6% faculty, and 9% other. Only 21% of the survey respondents had an academic qualification in STEM, which suggests that the sample is unlikely to be biased towards tech-savvy users.

### C. Summary

We collected data from February 5 to March 17, 2016. PhoneLab allows any participant to opt-out of an experiment at any time. Thus, of the 203 participants who installed our

<sup>2</sup>An Android `Activity` represents the application screen and UI elements currently exposed to the user.

custom Android build, there were 131 who used it for more than 20 days. During the study period, we collected 176M events across all participants (31K events per participant/day). Our dataset consists of 1,686 unique applications and 13K unique activities. Participants also responded to 4,636 prompts during the study period. We logged 96M sensitive permission requests, which translates to roughly one sensitive permission request every 6 seconds per participant. For the remainder of the paper, we only consider the data from the 131 participants who used the system for at least 20 days, which corresponds to 4,224 ESM prompts.

Of the 4,224 prompts, 55.3% were in response to `ACCESS_WIFI_STATE`, when trying to access WiFi SSID information that could be used to infer the location of the smartphone; 21.0%, 17.3%, 5.08%, 0.78%, and 0.54% were from accessing location directly, reading SMS, sending SMS, reading call logs, and accessing browser history, respectively. A total of 137 unique applications triggered prompts during the study period. Of the 4,224 prompts, participants wanted to deny 60.01% of them, and 57.65% of the prompts were shown when the requesting application was running in the foreground or the user had visual cues that the application was running (e.g., notifications). A Wilcoxon signed rank test with continuity correction revealed a statistically significant difference in participants’ desire to allow or deny a permission request based on the visibility of the requesting application ( $p < 0.0152$ ,  $r = 0.221$ ), which corroborates previous findings [43].

## IV. TYPES OF USERS

We hypothesized that there may be different types of users based on how they want to disclose their private information to third parties. It is imperative to identify these different sub-populations since different permission models affect users differently based on their privacy preferences; performance numbers averaged across a user population could be misleading since different sub-populations might react differently to the same permission model.

While our study size was too small to effectively apply clustering techniques to generate classes of users, we did find a meaningful distinction using the denial rate (i.e., the percentage of prompts to which users wanted to deny access). We aggregated users by their denial rate in 10% increments and examined how these different participants considered the surrounding contextual circumstances in their decisions.

We discovered that application visibility was a significant factor for users with a denial rate of 10–90%, but not for users with a denial rate of 0–10% or 90–100%. We call the former group *Contextuals*, as they seem to care about the surrounding context (i.e., they make nuanced decisions, allowing or denying a permission request based on whether they had contextual cues that indicated that the requesting application was running), and the latter group *Defaulters*, because they seem to simply always allow or always deny requests, regardless of contextual cues.

*Defaulters* accounted for 53% of 131 participants and *Contextuals* accounted for 47%. A Wilcoxon signed-rank test with

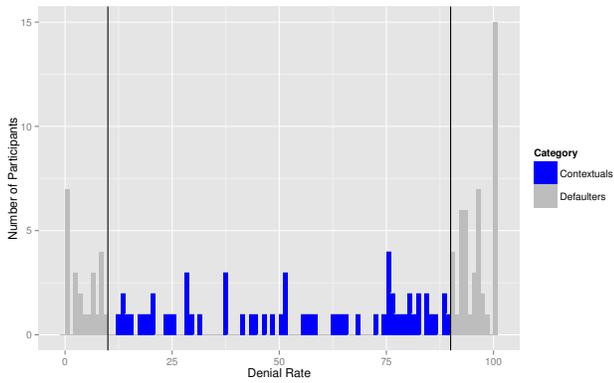


Fig. 2. Histogram of users based on their denial rate. *Defaulters* tended to allow or deny almost all requests without regard for contextual cues, whereas *Contextuals* considered the visibility of the requesting application.

Policy	Contextuals	Defaulters	Overall	Prompts
AOI	44.11%	6.00%	25.00%	0.00
<b>AOFU-AP</b>	<b>64.49%</b>	<b>93.33%</b>	<b>84.61%</b>	<b>12.34</b>
AOFU-APV	64.28%	92.85%	83.33%	15.79
AOFU- $A_F$ PV	66.67%	98.95%	84.61%	16.91
AOFU-VP	58.65%	94.44%	78.04%	6.43
AOFU-VA	63.39%	93.75%	84.21%	12.24
AOFU-A	64.27%	93.54%	83.33%	9.06
AOFU-P	57.95%	95.45%	82.14%	3.84
AOFU-V	52.27%	95.34%	81.48%	2.00

TABLE III

THE ACCURACY AND NUMBER OF DIFFERENT POSSIBLE ASK-ON-FIRST-USE COMBINATIONS. A: APPLICATION REQUESTING THE PERMISSION, P: PERMISSION TYPE REQUESTED, V: VISIBILITY OF THE APPLICATION REQUESTING THE PERMISSION,  $A_F$ : APPLICATION RUNNING IN THE FOREGROUND WHEN THE REQUEST IS MADE. AOFU-AP IS THE POLICY USED IN ANDROID MARSHMALLOW I.E., ASKING (PROMPTING) THE USER FOR EACH UNIQUE APPLICATION, PERMISSION COMBINATION. THE TABLE ALSO DIFFERENTIATES POLICY NUMBERS BASED ON THE SUBPOPULATION OF *Contextuals*, *Defaulters*, AND ACROSS ALL USERS.

continuity correction revealed a statistically significant difference in *Contextuals*' responses based on requesting application visibility ( $p < 0.013$ ,  $r = 0.312$ ), while for *Defaulters* there was no statistically significant difference ( $p = 0.227$ ). That is, *Contextuals* used visibility as a contextual cue, when deciding the appropriateness of a given permission request, whereas *Defaulters* did not vary their decisions based on this cue. Figure 2 shows the distribution of users based on their denial rate. Vertical lines indicate the borders between *Contextuals* and *Defaulters*.

In the remainder of the paper, we use our *Contextuals-Defaulters* categorization to measure how current and proposed models affect these two sub-populations, issues unique to these sub-populations, and ways to address these issues.

## V. ASK-ON-FIRST-USE PERMISSIONS

Ask-on-first-use (AOFU) is the current Android permission model, which was first adopted in Android 6.0 (Marshmallow). AOFU prompts the user whenever an application requests a *dangerous* permission for the first time [9]; the user's response

to this prompt is thereafter applied whenever the same application requests the same permission. As of March 2017, only 34.1% of Android users have Android Marshmallow or a higher version [10], and among these Marshmallow users, those who upgraded from a previous version only see runtime permission prompts for freshly-installed applications.

For the remaining 65.9% of users, the system policy is ask-on-install (AOI), which automatically allows all runtime permission requests. During the study period, all of our participants had AOI running as the default permission model. Because all runtime permission requests are allowed in AOI, any of our ESM prompts that the user wanted to deny correspond to mispredictions under the AOI model (i.e., the AOI model granted access to the data against users' actual preferences). Table III shows the expected median accuracy for AOI, as well as several other possible variants that we discuss in this section. The low median accuracy for *Defaulters* was due to the significant number of people who simply denied most of the prompts. The prompt count is zero for AOI because it does not prompt the user during runtime; users are only shown permission prompts at installation.

More users will have AOFU in the future, as they upgrade to Android 6.0 and beyond. To the best of our knowledge, no prior work has looked into quantifying the effectiveness of AOFU systematically; this section presents analysis of AOFU based on prompt responses collected from participants and creates a baseline against which to measure our system's improvement. We simulate how AOFU performs through our ESM prompt responses. Because AOFU is deterministic, each user's response to the first prompt for each *application:permission* combination tells us how the AOFU model would respond for subsequent requests by that same combination. For participants who responded to more than one prompt for each combination, we can quantify how often AOFU would have been correct for subsequent requests. Similarly, we also measure the accuracy for other possible policies that the platform could use to decide whether to prompt the user. For example, the status quo is for the platform to prompt the user for each new *application:permission* combination, but how would accuracy (and the number of prompts shown) change if the policy were to prompt on all new combinations of *application:permission:visibility*?

Table III shows the expected median accuracy<sup>3</sup> for each policy based on participants' responses. For each policy, *A* represents the application requesting the permission, *P* represents the requested permission, *V* represents the visibility of the requesting application, and  $A_F$  represents the application running in the foreground when a sensitive permission request was made. For instance, AOFU-AP is the policy where the user will be prompted for each new instance of an *application:permission* combination, which the Android 6.0 model employs. The last column shows the number of runtime prompts a participant would see under each policy over the duration of the study, if that policy were to be

<sup>3</sup>The presented numbers—except for average prompt count, which was normally distributed—are median values, because the distributions were skewed.

implemented. Both AOFU-AP and AOFU- $A_{FPV}$  show about a  $4.9\times$  reduction in error rate compared to AOI; AOFU- $A_{FPV}$  would require more prompts over AOFU-AP, though yields a similar overall accuracy rate.<sup>4</sup> Moving forward, we focus our analysis only on AOFU-AP (i.e., the current standard).

Instances where the user wants to deny a permission and the policy instead allows it (false positives) are *privacy violations*, because they expose more information to the application than the user desires. Instances where the user wants to allow a permission, but the policy denies it (false negatives) are *functionality losses*. This is because the application is likely to lose some functionality that the user desired when it is incorrectly denied a permission. Privacy violations and functionality losses were approximately evenly split between the two categories for AOFU-AP: median privacy violations and median functionality losses were 6.6% and 5.0%, respectively.

The AOFU policy works well for *Defaulters* because, by definition, they tend to be consistent after their initial responses for each combination. In contrast, the decisions of *Contextuals* vary due to other factors beyond just the requesting application and the requested permission type. Hence, the accuracy of AOFU for *Contextuals* is significantly lower than the accuracy for *Defaulters*. This distinction shows that learning privacy preferences for a significant portion of users requires a deeper understanding of factors affecting their decisions, such as behavioral tendencies and contextual cues. As Table III suggests, superficially adding more contextual variables (such as visibility of the requesting application) does not necessarily help to increase the accuracy of the AOFU policy.

The context in which users are prompted under AOFU might be a factor affecting its ability to predict subsequent instances. In previous work [43], we found that the visibility of the requesting application is a strong contextual cue users use to vary their decisions. During the study period, under the AOFU-AP policy, 60% of the prompts could have occurred when the requesting application was visible to the participant—these prompts had an accuracy of 83.3% in predicting subsequent instances. In instances where participants were prompted when the requesting application was running invisibly to the user, AOFU-AP had an accuracy of 93.7% in predicting subsequent instances. A Wilcoxon signed-ranks test, however, did not reveal a statistically significant difference ( $p < 0.3735$ ).

Our estimated accuracy numbers for AOFU may be inflated because AOFU in deployment (Android 6 and above) does not filter permission requests that do not reveal any sensitive information. For example, an application can request the ACCESS\_FINE\_LOCATION permission to check whether the phone has a specific location provider, which does not leak sensitive information. Our AOFU simulation uses the invoked function to determine if sensitive data was *actually* accessed, and only prompts in those cases (in the interest of avoiding any false positives), a distinction that AOFU in Android does not make. Thus, an Android user would see a permission request

<sup>4</sup>While AOFU- $A_{FPV}$  has greater *median* accuracy when examining *Defaulters* and *Contextuals* separately, because the distributions are skewed, the median overall accuracy is identical to AOFU-AP when combining the groups.

prompt when the application examines the list of location providers, and if the permission is granted, would not subsequently see prompts when location data is actually captured. Previous work found that 79% of first-time permission requests do not reveal any sensitive information [43], and nearly 33.9% of applications that request these sensitive permission types do not access sensitive data at all. The majority of AOFU prompts in Marshmallow are therefore effectively false positives, which incorrectly serve as the basis for future decisions. Given this, AOFU’s average accuracy is likely less than the numbers presented in Table III. We therefore consider our estimates of AOFU to be an upper bound.

## VI. LEARNING PRIVACY PREFERENCES

Table III shows that a significant portion of users (the 47% classified as *Contextuals*) make privacy decisions that depend on factors other than the application requesting the permission, the permission requested, and the visibility of the requesting application. To make decisions on behalf of the user, we must understand what other factors affect their privacy decisions. We built a machine learning model trained and tested on our labeled dataset of 4,224 prompts collected from 131 users over the period of 42 days. This approach is equivalent to training a model based on runtime prompts from hundreds of users and using it to predict those users’ future decisions.

We focus the scope of this work by making the following assumptions. We assume that the platform, i.e., the Android OS, is trusted to manage and enforce permissions for applications. We assume that applications must go through the platform’s permission system to gain access to protected resources. We assume that we are in a non-adversarial machine-learning setting wherein the adversary does not attempt to circumvent the machine-learned classifier by exploiting knowledge of its decision-making process—though we do present a discussion of this problem and potential solutions in Section IX.

### A. Feature Selection

Using the behavioral, contextual, and aggregate features shown in Table II, we constructed 16K candidate features, formed by combinations of specific applications and actions. We then selected 20 features by measuring Gini importance through random forests [30], significance testing for correlations, and singular value decomposition (SVD). SVD was particularly helpful to address the sparsity and high dimensionality issues caused by features generated based on application and activity usage. Table IV lists the 20 features used in the rest of this work.

The behavioral features ( $B$ ) that proved predictive relate to browsing habits, audio/call traits, and locking behavior. All behavioral features were normalized per day/user and were scaled in the actual model. Features relating to browsing habits included the number of websites visited, the proportion of HTTPS-secured links visited, the number of downloads, and proportion of sites visited that requested location access. Features relating to locking behavior included whether users employed a passcode/PIN/pattern, the frequency of screen

Feature Group	Feature	Type
Behavioral Features (B)	Number of times a website is loaded to the Chrome browser.	Numerical
	Out of all visited websites, the proportion of HTTPS-secured websites.	Numerical
	The number of downloads through Chrome.	Numerical
	Proportion of websites requested location through Chrome.	Numerical
	Number of times PIN/Password was used to unlock the screen.	Numerical
	Amount of time spent unlocking the screen.	Numerical
	Proportion of times screen was timed out instead of pressing the lock button.	Numerical
	Frequency of audio calls.	Numerical
	Amount of time spent on audio calls.	Numerical
	Proportion of time spent on silent mode.	Numerical
Runtime Features (R1)	Application visibility (True/False)	Categorical
	Permission type	Categorical
	User ID	Categorical
	Time of day of permission request	Numerical
Aggregated Features (A)	Average denial rate for (A1) application:permission:visibility	Numerical
	Average denial rate for (A2) application <sub>F</sub> :permission:visibility	Numerical

TABLE IV

THE COMPLETE LIST OF FEATURES USED IN THE ML MODEL EVALUATION. ALL THE NUMERICAL VALUES IN THE BEHAVIORAL GROUP ARE NORMALIZED PER DAY. WE USE ONE-HOT ENCODING FOR CATEGORICAL VARIABLES. WE NORMALIZED NUMERICAL VARIABLES BY MAKING EACH ONE A Z-SCORE RELATIVE TO ITS OWN AVERAGE.

unlocking, the proportion of times they allowed the screen to timeout instead of pressing the lock button, and the average amount of time spent unlocking the screen. Features under the audio and call category were the frequency of audio calls, the amount of time they spend on audio calls, and the proportion of time they spent on silent mode.

Our runtime features ( $R1/R2$ ) include the requesting application’s visibility, permission requested, and time of day of the request. Initially, we included the user ID to account for user-to-user variance, but as we discuss later, we subsequently removed it. Surprisingly, the application requesting the permission was not predictive, nor were other features based on the requesting application, such as application popularity.

Different users may have different ways of perceiving privacy threats posed by the same permission request. To account for this, the learning algorithm should be able to determine how each user perceives the appropriateness of a given request in order to accurately predict future decisions. To quantify the difference between users in how they perceive the threat posed by the same set of permission requests, we introduced a set of *aggregate features* that could be measured at runtime and that may partly capture users’ privacy preferences. We compute the average denial rate for each unique combination of *application:permission:visibility* (A1) and of *application<sub>F</sub>:permission:visibility* (A2). These aggregate features indicate how the user responded to previous prompts associated with that combination. As expected, after

<sup>5</sup>The application running in the foreground when the permission is requested by another application.

Feature Set	Contextuals	Defaulters	Overall
R1	69.30%	95.80%	83.71%
R2 + B	69.48%	95.92%	83.93%
R2 + A	75.45%	99.20%	92.24%

TABLE V

THE MEDIAN ACCURACY OF THE MACHINE LEARNING MODEL FOR DIFFERENT FEATURE GROUPS ACROSS DIFFERENT SUB POPULATIONS.

we introduced the aggregate features, the relative importance of the user ID variable diminished and so we removed it (i.e., users no longer needed to be uniquely identified). We define  $R2$  as  $R1$  without the user ID.

### B. Inference Based on Behavior

One of our main hypotheses is that passively observing users’ behaviors helps infer users’ future privacy decisions. To this end, we instrumented Android to collect a wide array of behavioral data, listed in Table II. We categorize our behavioral instrumentation into interaction with Android privacy/security settings, locking behavior, audio settings and call habits, web-browsing habits, and application usage habits. After the feature selection process (§VI-A), we found that only locking behavior, audio habits, and web-browsing habits correlated with privacy behaviors. Appendix B contains more information on feature importance. All the numerical values under the behavioral group were normalized per day.

We trained an SVM model with an RBF kernel on only the behavioral and runtime features listed in Table IV, excluding user ID. The 5-fold cross-validation accuracy (with random splitting) was 83% across all users. This first setup assumes we have prior knowledge of previous privacy decisions to a certain extent from each user before inferring their future privacy decisions, so it is primarily relevant after the user has been using their phone for a while. However, the biggest advantage of using behavioral data is that it can be observed passively without any active user involvement (i.e., no prompting).

We use leave-one-out cross validation to measure the extent to which we can infer user privacy decisions with *absolutely no user involvement* (and without any prior data on a user). In this second setup, when a new user starts using a smartphone, we assume there is a ML model which is already trained with behavioral data and privacy decisions collected from a selected set of other users. We then measured the efficacy of such a model to predict the privacy decisions of a new user, purely based on passively observed behavior and runtime information on the request, without ever prompting that new user. This is an even stricter lower bound on user involvement, which essentially mandates that a user has to make no effort to indicate privacy preferences, something that no system currently does.

We performed leave-one-out cross validation for each of our 131 participants, meaning we predicted a single user’s privacy decisions using a model trained using the data from the other 130 users’ privacy decisions and behavioral data. The only input for each test user was the passively observed

behavioral data and runtime data surrounding each request. The model yielded a median accuracy of 75%, which is a 3× improvement over AOI. Furthermore, AOI requires users to make active decisions during the installation of an application, which our second model does not require.

Examining only behavioral data with leave-one-group-out cross validation yielded a median accuracy of 56% for *Contextuals*, while for *Defaulters* it was 93.01%. Although, prediction using solely behavioral data fell short of AOFU-AP for *Contextuals*, it yielded a similar median accuracy for *Defaulters*; AOFU-AP required 12 prompts to reach this level of accuracy, whereas our model would not have resulted in any prompts. This relative success presents the significant observation that behavioral features, observed passively without user involvement, are useful in learning user privacy preferences. This provides the potential to open entirely new avenues of user learning and reduce the risk of habituation.

### C. Inference Based on Contextual Cues

Our SVM model with an RBF kernel produced the best accuracy. The results in the remainder of this section are trained and tested with five-fold cross validation with random splitting for an SVM model with an RBF kernel using the *ksvm* library in R. In all instances, the training set was bootstrapped with an equal number of allow and deny data points to avoid training a biased model. For each feature group, all hyperparameters were tuned through grid search to achieve highest accuracy. We used one-hot encoding for categorical variables. We normalized numerical variables by making each one a z-score relative to its own average. Table V shows how the median accuracy changes with different feature groups. As a minor note, the addition of the mentioned behavioral features to runtime features performed only marginally better; this could be due to the fact that those two groups do not complement each other in predictions. In this setup, we assume that there is a single model across all the users of Android.

By incorporating user involvement in the form of prompts, we can use our aggregate features to increase the accuracy for *Contextuals*, slightly less so for *Defaulters*. The aggregate features primarily capture how consistent users are for particular combinations (i.e., *application:permission:visibility*, *application<sub>F</sub>:permission:visibility*), which greatly affects accuracy for *Contextuals*. *Defaulters* have high accuracy with just runtime features (*R1*), as they are likely to stick with a default allow or deny policy regardless of the context surrounding a permission. Thus, even without any aggregate features (which do not impart any new information about this type of user), the model can predict privacy preferences of *Defaulters* with a high degree of accuracy. On the other hand, *Contextuals* are more likely to vary their decision for a given permission request. However, as the accuracy numbers in Table V suggest, this variance is correlated with some contextual cues. The high predictive power of aggregate features indicates that they may be capturing the contextual cues, used by *Contextuals* to make decisions, to a greater extent.

The fact that both *application:permission:visibility* and *application<sub>F</sub>:permission:visibility* are highly predictive (Appendix A) indicates that user responses for these combinations are consistent. The high consistency could relate to the notion that the visibility and the foreground application (application<sub>F</sub><sup>6</sup>) are strong contextual cues people use to make their privacy decisions; the only previously studied contextual cue was the visibility of the application requesting the sensitive data [43]. We offer a hypothesis for why foreground application could be significant: the sensitivity of the foreground application (i.e., high-sensitivity applications like banking, low-sensitivity applications like games) might impact how users perceive threats posed by requests. Irrespective of the application requesting the data, users may be likely to deny the request because of the elevated sense of risk. We discuss this further in §IX.

The model trained on feature sets *R2*, *A1*, and *A2* had the best accuracy (and the fewest privacy violations). For the remainder of the paper, we will refer to this model unless otherwise noted. We now compare AOFU-AP (the status quo as of Android 6.0 and above, presented in Table III) and our model (Table V). Across all users, our model reduced the error rate from 15.38% to 7.76%, nearly a two-fold improvement.

Mispredictions (errors) in the ML model were split between privacy violations and functionality losses (54% and 46%). Deciding which error type is more acceptable is subjective and depends on factors like the usability issues surrounding functionality losses and gravity of privacy violations. However, the (approximately) even split between the two error types shows that the ML is not biased towards one particular decision (denying vs. allowing a request). Furthermore, the area under the ROC curve (AUC), a metric used to measure the fairness of a classifier, is also significantly better in the ML model (0.936 as opposed to 0.796 for AOFU). This indicates that the ML model is equally good at predicting when to both allow and deny a permission request, while AOFU tends to lean more towards one decision. In particular, with the AOFU policy, users would experience privacy violations for 10.01% of decisions, compared to just 4.2% with the ML model. Privacy violations are likely more costly to the user than functionality loss: denied data can always be granted at a later time, but disclosed data cannot be taken back.

While increasing the number of prompts improves classifier accuracy, it plateaus after reaching its maximum accuracy, at a point we call the *steady state*. For some users, the classifier might not be able to infer their privacy preferences effectively, regardless of the number of prompts. As a metric to measure the effectiveness of the ML model, we measure the confidence of the model in the decisions it makes, based on prediction class probabilities.<sup>7</sup> In cases where the confidence of the model

<sup>6</sup>Even when the requesting application is running visible to the user, the foreground application could still be different from the requesting application since the only visible cue of the requesting application could be a notification in the notification bar.

<sup>7</sup>To calculate the class probabilities, we used the KSVM library in R. It employs a technique proposed by Platt et al. [25] to produce a numerical value for each class’s probability.

is below a certain threshold, the system should use a runtime prompt to ask the user to make an explicit decision. Thus, we looked into the prevalence of low-confidence predictions among the current predictions. With a 95% confidence interval, on average across five folds, low-confidence predictions accounted for less than 10% of all predictions. The remaining high-confidence predictions (90% of all predictions) had an average accuracy of 96.2%, whereas predictions with low confidence were only predicted with an average accuracy of 72%. §VII-B goes into this aspect in detail and estimates the rate at which users will see prompts in steady state.

The caveat in our ML model is that AOFU-AP only resulted in 12 prompts on average per user during the study, while our model averaged 24. The increased prompting stems from multiple prompts for the same combination of *application:permission:visibility*, whereas in AOFU, prompts are shown only once for each *application:permission* combination. During the study period, users on average saw 2.28 prompts per unique combination. While multiple prompts per combination help the ML model to capture user preferences under different contextual circumstances, it risks habituation, which may eventually reduce the reliability of the user responses.

The evaluation setup mentioned in the current section does not have a specific strategy to select the training set. It randomly splits the data set into the 5 folds and picks 4 out of 5 as the training set. In a real-world setup, the platform needs a strategy to carefully select the training set so that the platform can learn most of the user’s privacy preferences with a minimum number of prompts. The next section presents an in-depth analysis on possible ways to reduce the number of prompts needed to train the ML model.

## VII. LEARNING STRATEGY

This sections presents a strategy the platform can follow in the learning phase of a new user. The key objective of the learning strategy should be to learn the user’s privacy preferences with minimal user involvement (prompts). Once the model reaches adequate training, we can use model decision confidence to analyze how the ML model performs for different users and examine the tradeoff between user involvement and accuracy. We also utilize the model’s confidence on decisions to present a strategy that can further reduce model error through selective permission prompting.

### A. Bootstrapping

The *bootstrapping* phase occurs when the ML model is presented with a new user about whom the model has no prior information. In this section, we analyze how the accuracy improves as we prompt the user. Since the model presented in §VI is a single model trained with data from all users, the ML model can still predict a new user’s privacy decisions by leveraging the data collected on other users’ preferences.

We measured the accuracy of the ML model as if it had to predict each user’s prompt responses using a model trained using other users’ data. Formally, this is called leave-one-out cross-validation, where we remove all the prompt responses

from a single user. The training set contains all the prompt responses from 130 users and the test set is the prompt responses collected from the single remaining user. The model had a median accuracy of 66.6% (56.2% for *Contextuals*, 86.4% for *Defaulters*). Although this approach does not prompt new users, it falls short of AOFU. This no-prompt model behaves close to random guessing for *Contextuals* and significantly better for *Defaulters*. Furthermore, Wijesekera et al. found that individuals’ privacy preferences varied a lot [43], suggesting that utilizing other users’ decisions to predict decisions for a new user has limited effectiveness, especially for *Contextuals*; some level of prompting is necessary.

There are a few interesting avenues to explore when determining the optimal way to prompt the user in the learning phase. One option would be to follow the same weighted-reservoir sampling algorithm mentioned in §III-A. The algorithm is weighted by the frequency of each *application:permission:visibility* combination. The most frequent combination will have the highest probability of creating a permission prompt and after the given combination reaches a maximum of three prompts, the algorithm will no longer consider that combination for prompting, giving the second most frequent combination the new highest probability. Due to frequency-weighting and multiple prompts per combination, the weighted-reservoir sampling approach requires more prompts to cover a broader set of combinations. However, AOFU prompts only once per combination without frequency-weighting. This may be a useful strategy initially for a new user since it allows the platform to learn about the users’ privacy preferences for a wide array of combinations with minimal user interaction.

To simulate such an approach, we extend the aforementioned no-prompt model (leave-one-out validation). In the no-prompt model, there was no overlap of users in the train and test set. In the new approach, the training set includes the data from other users as well as the new user’s responses to the first occurrence of each unique combination of *application:permission:visibility*. The first occurrence of each unique combination simulates the AOFU-APV policy. That is, this model is bootstrapped using data from other users and then adopts the AOFU-APV policy to further learn the current user’s preferences. The experiment was conducted using the same set of features mentioned in §VI-A ( $R2 + A1 + A2$  and an SVM with a RBF kernel). The test set only contained prompt responses collected after the last AOFU prompt to ensure chronological consistency.

Figure 3 shows how accuracy changes with the varying number of AOFU prompts for *Contextuals* and *Defaulters*. For each of the 131 users, we ran the experiment varying the AOFU prompts from 1 to 12. We chose this upper bound because, on average, a participant saw 12 different unique *application:permission* combinations during the study period—the current permission model in Android. AOFU relies on user prompts for each new combination. The proposed ML model, however, has the advantage of leveraging data collected from other users to predict a combination not seen by the user; it can

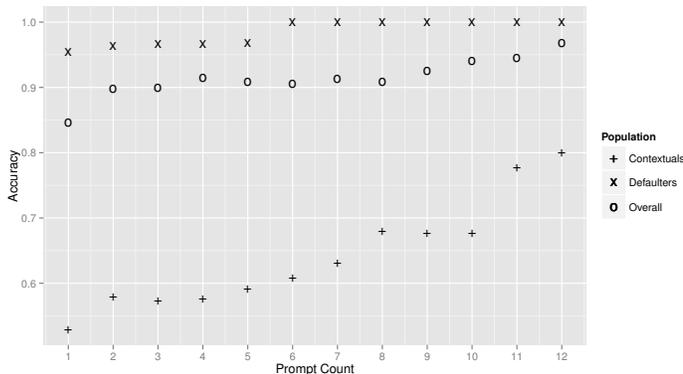


Fig. 3. How the median accuracy varies with the number of seen prompts

significantly reduce user involvement in the learning phase. After 12 prompts, accuracy reached 96.8% across all users.

Each new user starts off with a single model shared by all new users and then moves onto a separate model trained with AOFU prompt responses. We analyze its performance for *Defaulters* and *Contextuals* separately, finding that it improves accuracy while reducing user involvement in both cases, compared to the status quo.

We first examine how our model performs for *Defaulters*, 53% of our sample. Figure 3 shows that our model trained with AOFU permission-prompt responses outperforms AOFU from the very beginning. The model starts off with 96.6% accuracy (before it reaches close to 100% after 6 prompts), handily exceeding AOFU’s 93.33%. This is a 83.3% reduction in permission prompts compared to AOFU-AP (the status quo). Even with such a significant reduction in user involvement, the new approach cuts the prediction error rate in half.

*Contextuals* needed more prompts to outperform the AOFU policy; the hybrid approach matches AOFU-AP with just 7 prompts, a 42% reduction in prompts. With 12 permission prompts, same as needed for AOFU-AP, the new approach had reduced the error rate by 43% over AOFU-AP (the status quo). The number of prompts needed to reach this level of accuracy in the new approach is 25% less than what is needed for AOFU-APV. We also observed that as the number of prompts increased, the AUC of our predictions also similarly increased. Overall, the proposed learning strategy reduced the error rate by 80% after 12 user prompts over AOFU-AP. Given, *Defaulters* plateau early in their learning cycle (after only 6 prompts), the proposed learning strategy, on average, needs 9 prompts to reach its maximum capacity, which is a 25% reduction in user involvement over AOFU-AP.

*Contextuals* have a higher need for user involvement than *Defaulters*, primarily because it is easy to learn about *Defaulters*, as they are more likely to be consistent with early decisions. On the other hand, *Contextuals* vary their decisions based on different contextual cues and require more user involvement for the model to learn the cues used by each user and how do they affect their decisions. Thus, it is important to

find a way to differentiate between *Defaulters* and *Contextuals* early in the bootstrapping phase to determine which users require fewer prompts. The analysis of our hybrid approach addresses the concern of a high number of permission prompts initially for an ML approach. Over time, accuracy can always be improved with more prompts.

Our new hybrid approach of using AOFU-style permission prompts in the bootstrapping phase to train our model can achieve higher accuracy than AOFU, with significantly fewer prompts. Having a learning strategy (use of AOFU) over random selection helped to minimize user involvement (24 vs. 9) while significantly reducing the error rate (7.6% vs. 3.2%) over a random selection of the training set.

### B. Decision Confidence

In the previous section, we looked into how we can optimize the learning phase by merging AOFU and the ML model to reach higher accuracy with minimal user prompts. However, for a small set of users, more permission prompts will not increase accuracy, regardless of user involvement in the bootstrapping phase. This could be due to the fact that a portion of users in our dataset are making random decisions, or that the features that our ML model takes into account are not predictive of those users’ decision processes. While we do not have the data to support either explanation, we examine how we can measure whether the ML model will perform well for a particular user and quantify how often it does not. We present a method to identify difficult-to-predict users and reduce permission prompting for those users.

While running the experiment in §VII-A, we also measured how confident the ML model was for each decision it made. To measure the ML model’s confidence, we record the probability for each decision; since it is a binary classification (deny or allow), the closer the probability is to 0.5, the less confident it is. We then chose a *class probability threshold* above which a decision would be considered a high-confidence decision. In our analysis, we choose a class probability threshold of 0.6, since this value resulted in >96% accuracy for our fully-trained model ( $\approx 25$  prompts per user) for high-confidence decisions, but this is a tunable threshold. Thus, in the remainder of our analysis, decisions that the ML model made with a probability of >0.60 were labeled as high-confidence decisions, while those made with a probability of <0.60 were labeled as low-confidence decisions.

Since the most accurate version of AOFU uses 12 prompts, we also evaluate the confidence of our model after 12 AOFU-style prompts. This setup is identical to the bootstrapping approach; the model we evaluate here is trained on responses from other users and the first 12 prompts chosen by AOFU. With this scheme, we found that 10 users (7.63% of 131 users) had at least one decision predicted with low confidence. The remaining 92.37% of users had all privacy decisions predicted with high confidence. Among those users whose decisions were predicted with low confidence, the proportion of low-confidence decisions on average accounted for 17.63% (median = 16.67%) out of all their predicted decisions. With

a sensitive permission request once every 15 seconds [43], prompting even for 17.63% of predictions is not practical. Users who had low-confidence predictions had a median accuracy of 60.17%, compared to 98% accuracy for the remaining set of users with only high-confidence predictions. Out of the 10 users who had low-confidence predictions, there were no *Defaulters*. This further supports the observation in Figure 3 that *Defaulters* require a shorter learning period.

In a real-world scenario, after the platform (ML model) prompts the user for the first 12 AOFU prompts, the platform can measure the confidence of predicting unlabeled data (sensitive permission requests for which the platform did not prompt the user). If the proportion of low-confidence predictions is below some threshold, the ML model can be deemed to have successfully learned user privacy preferences and the platform should keep on using the regular permission-prompting strategy. Otherwise, the platform may choose to limit prompts (i.e., two per unique *application:permission:visibility* combination). It should also be noted that rather than having a fixed number of prompts (e.g., 12) to measure the low-confidence proportion, the platform can keep track of the low-confidence proportion as it prompts the user according to any heuristic (i.e., unique combinations). If the proportion does not decrease with the number of prompts, we can infer that the ML model is not learning user preferences effectively or the user is making random decisions, indicating that limiting prompts and accepting lower accuracy could be a better option for that specific user, to avoid excessive prompting. However, depending on which group the user is in (*Contextual* or *Defaulter*), the point at which the platform could make the decision to continue or limit prompting could change. In general, the platform should be able to reach this deciding point relatively quickly for *Defaulters*.

Among participants with no low-confidence predictions, we had a median error rate of 2% (using the new hybrid approach after 12 AOFU prompts); for the same set of users, AOFU could only reach a median error rate of 13.3%. However, using AOFU, a user in that set would have needed an average of 15.11 prompts to reach that accuracy. Using the ML model, a user would need just 9 prompts on average (*Defaulters* require far fewer prompts, dropping the average); the model only requires 60% of the prompts that AOFU requires. Even with far fewer prompts in the learning phase, the ML model achieves a 84.61% reduction in error rate relative to AOFU.

While our model may not perform well for all users, it does seem to work quite well for the majority of users (92.37% of our sample). We provide a way of quickly identifying users for whom our system does not perform well, and propose limiting prompts to avoid excessive user burden for those users, at the cost of reduced efficacy. In the worst case, we could simply employ the AOFU model for users our system does not work well for, resulting in a multifaceted approach that is at least as good as the status quo for all users.

### C. Online Model

Our proposed system relies on training models on a trusted server, sending it to client phones (i.e., as a weight vector), and having phones make classifications. By utilizing an online learning model, we can train models incrementally as users respond to prompts over time. There are two key advantages to this: (i) this model adapts to changing user preferences over time; (ii) it distributes the overhead of training increasing the practicality of locally training the classifier on the phone itself.

Our scheme requires two components: a feature extraction and storage mechanism on the phone (a small extension to our existing instrumentation) and a machine learning pipeline on a trusted server. The phone sends feature vectors to the server every few prompts, and the server responds with a weight vector representing the newly trained classifier. To bootstrap the process, the server’s models can be initialized with a model trained on a few hundred users, such as our single model across all users. Since each user contributes data points over time, the online model adapts to changing privacy preferences even if they conflict with previous data. When using this scheme, each model takes less than 10 KB to store. With our current model, each feature and weight vector are at most 3 KB each, resulting in at most 6 KB of data transfer per day.

To evaluate the accuracy of our online model, we trained a classifier using stochastic gradient descent (SGD) with five-fold cross validation on our 4,224-point data set. This served as the bootstrapping phase. We then simulated receiving the remaining data one-at-a-time in timestamp order. Any features that changed with time (e.g., running averages for aggregate features, event counts) were computed with each incoming data point, creating a snapshot of features as the phone would see it. We then tested accuracy on the chronologically last 20% of our dataset. Our SGD classifier had 93.8% accuracy (AUC=0.929). We attribute the drop in accuracy (compared to our offline model) to the fact that running averages take multiple data points to reach steady-state, causing some earlier predictions to be incorrect.

A natural concern with a trusted server is compromise. To address this concern, we do not send any personally-identifiable data to the server, and any features sent to the server are *scaled*; they are reported in standard deviations from the mean, not in raw values. Furthermore, using an online model with incremental training allows us to periodically train the model on the phone (i.e., nightly, when the user is charging her device) to eliminate the need for a trusted server.

## VIII. CONTEXTUAL INTEGRITY

Contextual integrity is a conceptual framework that helps explain why most permission models fail to protect user privacy—they often do not take the context surrounding privacy decisions into account. In addressing this issue, we propose an ML model that infers when context has changed. We believe that this is an important first step towards operationalizing the notion of *contextual integrity*. In this section, we explain the observations that we made in §VI-C based on the contextual integrity framework proposed by Barth et al. [6].

Contextual integrity provides a conceptual framework to better understand how users make privacy decisions; we use Barth et al.’s formalized model [6] as a framework in which to view Android permission models. Barth et al. model parties as communicating agents ( $P$ ) knowing information represented as attributes ( $T$ ). A knowledge state  $\kappa$  is defined as a subset of  $P \times P \times T$ . We use  $\kappa = (p, q, t)$  to mean that agent  $p$  knows attribute  $t$  of agent  $q$ . Agents play roles ( $R$ ) in contexts ( $C$ ).

For example, an agent can be a game application, and have the role of a game provider in an entertainment context. Knowledge transfer happens when information is communicated between agents; all communications can be represented through a series of traces  $(\kappa, (p, r), a)$ , which are combinations of a knowledge state  $\kappa$ , a role state  $(p, r)$ , and a communication action  $a$  (information sent). The role an agent plays in a given context helps determine whether an information flow is acceptable for a user. The relationship between the agent sending the information and the role of the agent ( $(p, r)$ ) receiving the information must follow these contextual norms.

With the Android permission model, the same framework can be applied. Both the user and the third-party application are communicating agents, and the information to be transferred is the sensitive data requested by the application. When a third-party application requests permission to access a guarded resource (e.g., location), knowledge of the guarded resource is transferred from the one agent (i.e., the user/platform) to another agent (i.e., the third-party application). The extent to which a user expects a given request depends not on the agent (the application requesting the data), but on the role that agent is playing in that context. This explains why the application as a feature itself (i.e., application name) was not predictive in our models: this feature does not represent the role when determining whether it is unexpected. While it is difficult for the platform to determine the exact role an application is playing, the visibility of the application hints at its role. For instance, when the user is using Google Maps to navigate, it is playing a different role from when Google Maps is running in the background without the user’s knowledge. We believe that this is the reason why the visibility of the requesting application is significant: it helps the user to infer the role played by the application requesting the permission.

The user expects applications in certain roles to access resources depending on the context in which the request is made. We believe that the foreground application sets this context. Thus a combination of the role and the context decides whether an information flow is expected to occur or not. Automatically inferring the exact context of a request is likely an intractable problem. For our purposes, however, it is possible that we need to only infer when context has *changed*, or rather, when data is being requested in a context that is no longer acceptable to the user. Based on our data, we believe that features based on foreground application and visibility are most useful for this purpose, from our collected dataset.

We now combine all of this into a concrete example within the contextual integrity framework: If a user is using Google Maps to reach a destination, the application can play the

role of a navigator in a geolocation context, whereby the user feels comfortable sharing her location. In contrast, if the same application requests location while running as a service invisible to the user, the user may not want to provide the same information. Background applications play the role of “passive listeners” in most contexts; this role as perceived by the user may be why background applications are likelier to violate privacy expectations and consequently be denied by users.

AOFU primarily focuses on controlling access through rules for *application:permission* combinations. Thus, AOFU neglects the role played by the application (visibility) and relies purely on the agent (the application) and the information subject (permission type). This explains why AOFU is wrong in nearly one-fifth of cases. Based on Table III, both AOFU-VA (possibly identifying the role played by the application) and AOFU- $A_F$ PV (possibly identifying the current context because of the current foreground application- $A_F$ ) have higher accuracy than the other AOFU combinations. However, as the contextual integrity framework suggests, the permission model has to take both the role and the current context into account before making an accurate decision. AOFU (and other models that neglect context) only makes it possible to consider a single aspect, a limitation that does not apply to our model.

While the data presented in this work suggest the importance of capturing context to better protect user privacy, more work is needed along these lines to fully understand how people use context to make decisions in the Android permission model. Nevertheless, we believe we contribute a significant initial step towards applying contextual integrity to improve smartphone privacy by dynamically regulating permissions.

## IX. DISCUSSION

The primary goal of this research was to improve the accuracy of the Android permission system so that it more correctly aligns with user privacy preferences. We began with four hypotheses: (i) that the currently deployed AOFU policy frequently violates user privacy; (ii) that the contextual information it ignores is useful; (iii) that a ML-based classifier can account for this contextual information and thus improve on the status quo; and (iv) that passively observable behavioral traits can be used to infer privacy preferences.

To test these hypotheses, we performed the first large-scale study on the effectiveness of AOFU permission systems in the wild, which showed that hypotheses (i) and (ii) hold. We further built an ML classifier that took user permission decisions along with observations of user behaviors and the context surrounding those decisions to show that (iii) and (iv) hold. Our results show that existing systems have significant room for improvement, and other permission-granting systems may benefit from applying our results.

### A. Limitations of Permission Models

Our field study confirms that users care about their privacy and are wary of permission requests that violate their expectations. We observed that 95% of participants chose to block at least one permission request; in fact, the average denial

rate was 60%—a staggering amount given that the AOI model permits all permission requests for an installed application.

While AOFU improves over the AOI model, it still violates user privacy around one in seven times, as users deviate from their initial responses to permission requests. This amount is significant because of the high frequency of sensitive permission requests: a 15% error rate yields thousands of privacy violations per user—based on the latest dataset, this amounts to a potential privacy violation every minute. It further shows that AOFU’s correctness assumption—that users make binary decisions based only on the *application:permission* combination—is incorrect. Users take a richer space of information into account when making decisions about permission requests.

### B. Our ML-Based Model

We show that ML techniques are effective at learning from both the user’s previous decisions and the current environmental context in order to predict whether to grant permissions on the user’s behalf. In fact, our techniques achieve better results than the methods currently deployed on millions of phones worldwide—while imposing significantly less user burden.

Our work incorporates elements of the surrounding context into a machine-learning model. This better approximates user decisions by finding factors relevant for users that are not encapsulated by the AOFU model. In fact, our ML model reduces the errors made by the AOFU model by 75%. Our ML model’s 97% accuracy is a substantial improvement over AOFU’s 85% and AOI’s 25%; the latter two of which comprise the *status quo* in the Android ecosystem.

Our research shows that many users make neither random nor fixed decisions: the environmental context plays a significant role in user decision-making. Automatically detecting the precise context surrounding a request for sensitive data is an incredibly difficult problem (e.g., inferring *how* data will be used), and is potentially intractable. However, to better support user privacy, that problem does not need to be solved; instead, we show that systems can be improved by using environmental data to infer when context has *changed*. We found that the most predictive factors in the environmental context were whether the application requesting the permission is visible, and what the foreground application the user is engaged with. These are both strong contextual cues used by users, insofar as they allowed us to better predict changes in context. Our results show that ML techniques have great potential in improving user privacy, by allowing us to infer when context has changed, and therefore when users would want data requests to be brought to their attention.

### C. Reducing the User Burden

Our work is also novel in using passively observable data to infer privacy decisions: we show that we can predict a user’s preferences without *any* permission prompts. Our model trained solely on behavioral traits yields a three-fold improvement over AOI; for *Defaulters*—who account for 53% of our sample—it was as accurate as AOFU-AP. These results demonstrate that we can match the status quo without *any*

active user involvement (i.e., the need for obtrusive prompts). These results imply that learning privacy preferences may be done entirely passively, which, to our knowledge, has not yet been attempted in this domain. Our behavioral feature set provides a promising new direction to guide research in creating permission models that minimize user burden.

The ML model trained with contextual data and past decisions also significantly reduced the user burden while achieving higher accuracy than AOFU. The model yielded an 81% reduction in prediction errors while reducing user involvement by 25%. The significance of this observation is that by reducing the risk of habituation, it increases reliability when user input is needed.

### D. User- and Permission-Tailored Models

Our ML-based model incorporates data from all users into a single predictive model. It may be the case, however, that a collection of models tailored to particular types of users outperforms our general-purpose model—provided that the correct model is used for the particular user and permission. To determine if this is true, we clustered users into groups based first on their behavioral features, and then their denial rate, to see if we could build superior cluster-tailored ML models. Having data for only 131 users, however, resulted in clusters too small to carry out an effective analysis. We note that we also created a separate model for each sensitive permission type, using data only for that permission. Our experiments determined, however, that these models were no better (and often worse) than our general model. It is possible that such tailored models may be more useful when our system is implemented at scale.

### E. Attacking the ML Model

Attacking the ML model to get access to users’ data without prompting is a legitimate concern [5]. There are multiple ways an adversary can influence the proposed permission model: (i) imposing an adversarial ML environment [31]; (ii) polluting the training set to bias the model to accept permissions; and (iii) manipulating input features in order to get access without user notification. We assume in this work that the platform is not compromised; a compromised platform will degrade any permission model’s ability to protect resources.

A thorough analysis on this topic is outside of our scope. Despite that, we looked at the possibility of manipulating features to get access to resources without user consent. None of the behavioral features used in the model can be influenced, since that would require compromising the platform. An adversary can control the runtime features for a given permission request by specifically choosing when to request the permission. We generated feature vectors manipulating every adversary-controlled value and combination from our dataset, and tested them on our model. We did not find any conclusive evidence that the adversary can exploit the ML model by manipulating the input features to get access to resources without user consent.

As this is not a comprehensive analysis on attack vectors, it is possible that a scenario exists where the adversary is able to access sensitive resources without prompting the user first. Our preliminary analysis suggests that such attacks may be non-trivial, but more work is needed to study and prevent such attacks, particularly examining adversarial ML techniques and feature brittleness.

#### F. Experimental Caveat

We repeat a caveat about our experimental data: users were free to deny permissions without any consequences. We explicitly informed participants in our study that their decisions to deny permission requests would have no impact on the actual behavior of their applications. This is important to note because if an application is denied a permission, it may exhibit undefined behavior or lose important functionality. In fact, researchers have noted that many applications crash when permissions are denied [13]. If these consequences are imposed on users, they may decide that the functionality is more important than their privacy decision.

If we actually denied permissions, users' decisions may skew towards a decreased denial rate. The denial rates in our experiments therefore represent the actual privacy preferences of users and their *expectations* of reasonable application behavior—not the result of choosing between application functionality and privacy. We believe that how people react when choosing between functionality and privacy preferences is an important research question beyond the scope of this paper. Such a change, however, will not limit this contribution, since our proposed model was effective in guarding resources of the users who are selective in their decision making—the proposed classifier reduced the error rate of *Contextuals* by 44%.

We believe that there are important unanswered questions about how to solve the technical hurdles surrounding enforcing restrictive preferences with minimal usability issues. As a first step towards building a platform that does not force users to choose between their privacy preferences and required functionality, we must develop an environment where permissions appear—to the application—to be allowed, but in reality only spurious or artificial data is provided.

#### G. Types of Users

We presented a categorization of users based on the significance that the application's visibility played towards their individual privacy decisions. We believe that in an actual permission denial setting, the distribution will be different from what was observed in our study. Our categorization's significance, however, motivates a deeper analysis on understanding the factors that divide *Contextuals* and *Defaulters*. While visibility was an important factor in this division, there may be others that are significant and relevant. More work needs to be done to explore how *Contextuals* make decisions and which behaviors correlate with their decisions.

#### H. User Interface Panel

Any model that predicts user decisions has the risk of making incorrect predictions. Making predictions on a user's

behalf, however, is necessary because permissions are requested by applications with too high a frequency for manual examination. While we do not expect any system to be able to obtain perfect accuracy, we do expect that our 97% accuracy can be improved upon.

One plausible way of improving the accuracy of the permission model is to empower the user to review and make changes on how the ML model makes decisions through a user feedback panel. This gives users recourse to correct undesirable decisions. The UI panel could also be used to reduce the usability issues and functionality loss stemming from permission denial. The panel should help the user figure out which rule incurred the functionality loss and to change it accordingly. A user may also use this to adjust their settings as their privacy preferences evolve over time.

#### I. The Cost of Greater Control

A more restrictive platform means users will have greater control over the data being shared with third parties. Applications that generate revenue based on user data, however, could be cut off from their primary revenue source. Such an effect could disrupt the current eco-system and force app developers to degrade app functionality based on the availability of the data. We believe the current eco-system is unfairly biased against users and tighter control will make the user an equal stakeholder. While more work is needed to understand the effects of a more restrictive platform, we believe it is imperative to let the user have greater control over their own data.

#### J. Conclusions

We have shown a number of important results. Users care about their privacy: they deny a significant number of requests to access sensitive data. Existing permission models for Android phones still result in significant privacy violations. Users may allow permissions some times, while denying them at others, implying that there are more factors that go into the decision-making process than simply the application name and the permission type. We collected real-world data from 131 users and found that application visibility and the current foreground application were important factors in user decisions. We used the data we collected to build a machine-learning model to make automatic permission decisions. One of our models had a comparable error rate to AOFU and benefited from not requiring any user prompting. Another of our models required some user prompts—less than is required by AOFU—and achieved a reduction of AOFU's error rate by 81%.

#### ACKNOWLEDGMENTS

This research was supported by the United States Department of Homeland Security's Science and Technology Directorate under contract FA8750-16-C-0140, the Center for Long-Term Cybersecurity (CLTC) at UC Berkeley, the National Science Foundation under grant CNS-1318680, and Intel through the ISTC for Secure Computing. The content of this document does not necessarily reflect the position or the policy of the U.S. Government and no official endorsement should be inferred.

## REFERENCES

- [1] Y. Agarwal and M. Hall, "Protectmyprivacy: Detecting and mitigating privacy leaks on ios devices using crowdsourcing," in *Proceeding of the 11th Annual International Conference on Mobile Systems, Applications, and Services*, ser. MobiSys '13. New York, NY, USA: ACM, 2013, pp. 97–110. [Online]. Available: <http://doi.acm.org/10.1145/2462456.2464460>
- [2] H. M. Almohri, D. D. Yao, and D. Kafura, "Droidbarrier: Know what is executing on your android," in *Proc. of the 4th ACM Conf. on Data and Application Security and Privacy*, ser. CODASPY '14. New York, NY, USA: ACM, 2014, pp. 257–264. [Online]. Available: <http://doi.acm.org/10.1145/2557547.2557571>
- [3] H. Almuhammedi, F. Schaub, N. Sadeh, I. Adjerid, A. Acquisti, J. Gluck, L. F. Cranor, and Y. Agarwal, "Your location has been shared 5,398 times!: A field study on mobile app privacy nudging," in *Proc. of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 2015, pp. 787–796.
- [4] K. W. Y. Au, Y. F. Zhou, Z. Huang, and D. Lie, "Pscout: Analyzing the android permission specification," in *Proc. of the 2012 ACM Conf. on Computer and Communications Security*, ser. CCS '12. New York, NY, USA: ACM, 2012, pp. 217–228. [Online]. Available: <http://doi.acm.org/10.1145/2382196.2382222>
- [5] M. Barreno, B. Nelson, R. Sears, A. D. Joseph, and J. D. Tygar, "Can machine learning be secure?" in *Proceedings of the 2006 ACM Symposium on Information, computer and communications security*. ACM, 2006, pp. 16–25.
- [6] A. Barth, A. Datta, J. C. Mitchell, and H. Nissenbaum, "Privacy and contextual integrity: Framework and applications," in *Proc. of the 2006 IEEE Symposium on Security and Privacy*, ser. SP '06. Washington, DC, USA: IEEE Computer Society, 2006. [Online]. Available: <http://dx.doi.org/10.1109/SP.2006.32>
- [7] I. Bilogrevic, K. Huguenin, B. Agir, M. Jadhwal, and J.-P. Hubaux, "Adaptive information-sharing for privacy-aware mobile social networks," in *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, ser. UbiComp '13. New York, NY, USA: ACM, 2013, pp. 657–666. [Online]. Available: <http://doi.acm.org/10.1145/2493432.2493510>
- [8] E. Bodden, "Easily instrumenting android applications for security purposes," in *Proc. of the ACM Conf. on Comp. and Comm. Sec.*, ser. CCS '13. NY, NY, USA: ACM, 2013, pp. 1499–1502. [Online]. Available: <http://doi.acm.org/10.1145/2508859.2516759>
- [9] A. Developer, "Requesting permissions," <https://developer.android.com/guide/topics/permissions/requesting.html>, accessed: March 18, 2017.
- [10] G. Developer, "Distribution of android versions," <http://developer.android.com/about/dashboards/index.html>, accessed: March 15, 2017.
- [11] S. Egelman, A. P. Felt, and D. Wagner, "Choice architecture and smartphone privacy: There's a price for that," in *The 2012 Workshop on the Economics of Information Security (WEIS)*, 2012.
- [12] W. Enck, P. Gilbert, 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," in *Proceedings of the 9th USENIX Conference on Operating Systems Design and Implementation*, ser. OSDI'10. Berkeley, CA, USA: USENIX Association, 2010, pp. 1–6. [Online]. Available: <http://dl.acm.org/citation.cfm?id=1924943.1924971>
- [13] Z. Fang, W. Han, D. Li, Z. Guo, D. Guo, X. S. Wang, Z. Qian, and H. Chen, "revdroid: Code analysis of the side effects after dynamic permission revocation of android apps," in *Proceedings of the 11th ACM Asia Conference on Computer and Communications Security (ASIACCS 2016)*. Xi'an, China: ACM, 2016.
- [14] A. P. Felt, E. Chin, S. Hanna, D. Song, and D. Wagner, "Android permissions demystified," in *Proc. of the ACM Conf. on Comp. and Comm. Sec.*, ser. CCS '11. New York, NY, USA: ACM, 2011, pp. 627–638. [Online]. Available: <http://doi.acm.org/10.1145/2046707.2046779>
- [15] A. P. Felt, S. Egelman, M. Finifter, D. Akhawe, and D. Wagner, "How to ask for permission," in *Proc. of the 7th USENIX conference on Hot Topics in Security*. Berkeley, CA, USA: USENIX Association, 2012. [Online]. Available: <http://dl.acm.org/citation.cfm?id=2372387.2372394>
- [16] A. P. Felt, S. Egelman, and D. Wagner, "I've got 99 problems, but vibration ain't one: a survey of smartphone users' concerns," in *Proc. of the 2nd ACM workshop on Security and Privacy in Smartphones and Mobile devices*, ser. SPSM '12. New York, NY, USA: ACM, 2012, pp. 33–44. [Online]. Available: <http://doi.acm.org/10.1145/2381934.2381943>
- [17] A. P. Felt, E. Ha, S. Egelman, A. Haney, E. Chin, and D. Wagner, "Android permissions: user attention, comprehension, and behavior," in *Proc. of the Eighth Symposium on Usable Privacy and Security*, ser. SOUPS '12. New York, NY, USA: ACM, 2012. [Online]. Available: <http://doi.acm.org/10.1145/2335356.2335360>
- [18] C. Gible, J. Crussell, J. Erickson, and H. Chen, "Androidleaks: Automatically detecting potential privacy leaks in android applications on a large scale," in *Proc. of the 5th Intl. Conf. on Trust and Trustworthy Computing*, ser. TRUST'12. Berlin, Heidelberg: Springer-Verlag, 2012, pp. 291–307. [Online]. Available: [http://dx.doi.org/10.1007/978-3-642-30921-2\\_17](http://dx.doi.org/10.1007/978-3-642-30921-2_17)
- [19] A. Gorla, I. Tavecchia, F. Gross, and A. Zeller, "Checking app behavior against app descriptions," in *Proceedings of the 36th International Conference on Software Engineering*, ser. ICSE 2014. New York, NY, USA: ACM, 2014, pp. 1025–1035. [Online]. Available: <http://doi.acm.org/10.1145/2568225.2568276>
- [20] S. E. Hormuth, "The sampling of experiences in situ," *Journal of personality*, vol. 54, no. 1, pp. 262–293, 1986.
- [21] P. Hornyack, S. Han, J. Jung, S. Schechter, and D. Wetherall, "These aren't the droids you're looking for: Retrofitting android to protect data from imperious applications," in *Proc. of the ACM Conf. on Comp. and Comm. Sec.*, ser. CCS '11. New York, NY, USA: ACM, 2011, pp. 639–652. [Online]. Available: <http://doi.acm.org/10.1145/2046707.2046780>
- [22] J. Jung, S. Han, and D. Wetherall, "Short paper: Enhancing mobile application permissions with runtime feedback and constraints," in *Proceedings of the Second ACM Workshop on Security and Privacy in Smartphones and Mobile Devices*, ser. SPSM '12. New York, NY, USA: ACM, 2012, pp. 45–50. [Online]. Available: <http://doi.acm.org/10.1145/2381934.2381944>
- [23] P. G. Kelley, S. Consolvo, L. F. Cranor, J. Jung, N. Sadeh, and D. Wetherall, "A conundrum of permissions: Installing applications on an android smartphone," in *Proc. of the 16th Intl. Conf. on Financial Cryptography and Data Sec.*, ser. FC'12. Berlin, Heidelberg: Springer-Verlag, 2012, pp. 68–79. [Online]. Available: [http://dx.doi.org/10.1007/978-3-642-34638-5\\_6](http://dx.doi.org/10.1007/978-3-642-34638-5_6)
- [24] W. Klieber, L. Flynn, A. Bhosale, L. Jia, and L. Bauer, "Android taint flow analysis for app sets," in *Proceedings of the 3rd ACM SIGPLAN International Workshop on the State of the Art in Java Program Analysis*, ser. SOAP '14, New York, NY, USA, 2014. [Online]. Available: <http://doi.acm.org/10.1145/2614628.2614633>
- [25] H.-T. Lin, C.-J. Lin, and R. C. Weng, "A note on platt's probabilistic outputs for support vector machines," *Machine learning*, vol. 68, no. 3, pp. 267–276, 2007.
- [26] J. Lin, B. Liu, N. Sadeh, and J. I. Hong, "Modeling users' mobile app privacy preferences: Restoring usability in a sea of permission settings," in *Symposium On Usable Privacy and Security (SOUPS 2014)*. Menlo Park, CA: USENIX Association, 2014, pp. 199–212. [Online]. Available: <https://www.usenix.org/conference/soups2014/proceedings/presentation/lin>
- [27] J. Lin, N. Sadeh, S. Amini, J. Lindqvist, J. I. Hong, and J. Zhang, "Expectation and purpose: understanding users' mental models of mobile app privacy through crowdsourcing," in *Proc. of the 2012 ACM Conf. on Ubiquitous Computing*, ser. UbiComp '12. New York, NY, USA: ACM, 2012, pp. 501–510. [Online]. Available: <http://doi.acm.org/10.1145/2370216.2370290>
- [28] B. Liu, M. S. Andersen, F. Schaub, H. Almuhammedi, S. A. Zhang, N. Sadeh, Y. Agarwal, and A. Acquisti, "Follow my recommendations: A personalized assistant for mobile app permissions," in *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*, 2016.
- [29] B. Liu, J. Lin, and N. Sadeh, "Reconciling mobile app privacy and usability on smartphones: Could user privacy profiles help?" in *Proceedings of the 23rd International Conference on World Wide Web*, ser. WWW '14. New York, NY, USA: ACM, 2014, pp. 201–212. [Online]. Available: <http://doi.acm.org/10.1145/2566486.2568035>
- [30] G. Louppe, L. Wehenkel, A. Sutera, and P. Geurts, "Understanding variable importances in forests of randomized trees," in *Advances in Neural Information Processing Systems 26*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2013. [Online]. Available: <http://papers.nips.cc/paper/4928-understanding-variable-importances-in-forests-of-randomized-trees.pdf>

- [31] D. Lowd and C. Meek, "Adversarial learning," in *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*. ACM, 2005, pp. 641–647.
- [32] K. Micinski, D. Votipka, R. Stevens, N. Kofinas, J. S. Foster, and M. L. Mazurek, "User interactions and permission use on android," in *CHI 2017*, 2017.
- [33] A. Nandugudi, A. Maiti, T. Ki, F. Bulut, M. Demirbas, T. Kosar, C. Qiao, S. Y. Ko, and G. Challen, "Phonelab: A large programmable smartphone testbed," in *Proceedings of First International Workshop on Sensing and Big Data Mining*. ACM, 2013, pp. 1–6.
- [34] H. Nissenbaum, "Privacy as contextual integrity," *Washington Law Review*, vol. 79, p. 119, February 2004.
- [35] T. Ringer, D. Grossman, and F. Roesner, "Audacious: User-driven access control with unmodified operating systems," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2016, pp. 204–216.
- [36] F. Roesner and T. Kohno, "Securing embedded user interfaces: Android and beyond," in *Presented as part of the 22nd USENIX Security Symposium (USENIX Security 13)*, 2013, pp. 97–112.
- [37] F. Roesner, T. Kohno, A. Moshchuk, B. Parno, H. J. Wang, and C. Cowan, "User-driven access control: Rethinking permission granting in modern operating systems," in *2012 IEEE Symposium on Security and Privacy*. IEEE, 2012, pp. 224–238.
- [38] J. L. B. L. N. Sadeh and J. I. Hong, "Modeling users' mobile app privacy preferences: Restoring usability in a sea of permission settings," in *Symposium on Usable Privacy and Security (SOUPS)*, 2014.
- [39] B. Shebaro, O. Oluwatimi, D. Midi, and E. Bertino, "Identidroid: Android can finally wear its anonymous suit," *Trans. Data Privacy*, vol. 7, no. 1, pp. 27–50, Apr. 2014. [Online]. Available: <http://dl.acm.org/citation.cfm?id=2612163,2612165>
- [40] M. Spreitzenbarth, F. Freiling, F. Echtler, T. Schreck, and J. Hoffmann, "Mobile-sandbox: Having a deeper look into android applications," in *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, ser. SAC '13. New York, NY, USA: ACM, 2013. [Online]. Available: <http://doi.acm.org/10.1145/2480362.2480701>
- [41] C. Thompson, M. Johnson, S. Egelman, D. Wagner, and J. King, "When it's better to ask forgiveness than get permission: Designing usable audit mechanisms for mobile permissions," in *Proc. of the 2013 Symposium on Usable Privacy and Security (SOUPS)*, 2013.
- [42] X. Wei, L. Gomez, I. Neamtii, and M. Faloutsos, "Permission evolution in the android ecosystem," in *Proceedings of the 28th Annual Computer Security Applications Conference*, ser. ACSAC '12. New York, NY, USA: ACM, 2012, pp. 31–40. [Online]. Available: <http://doi.acm.org/10.1145/2420950.2420956>
- [43] P. Wijesekera, A. Baokar, A. Hosseini, S. Egelman, D. Wagner, and K. Beznosov, "Android permissions remystified: A field study on contextual integrity," in *24th USENIX Security Symposium (USENIX Security 15)*. Washington, D.C.: USENIX Association, Aug. 2015, pp. 499–514. [Online]. Available: <https://www.usenix.org/conference/usenixsecurity15/technical-sessions/presentation/wijesekera>
- [44] H. Wu, B. P. Knijnenburg, and A. Kobsa, "Improving the prediction of users' disclosure behavior by making them disclose more predictably?" in *Symposium on Usable Privacy and Security (SOUPS)*, 2014.
- [45] K.-P. Yee, "Guidelines and strategies for secure interaction design," *Security and Usability: Designing Secure Systems That People Can Use*, vol. 247, 2005.
- [46] H. Zhu, H. Xiong, Y. Ge, and E. Chen, "Mobile app recommendations with security and privacy awareness," in *Proc. of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, 2014. [Online]. Available: <http://doi.acm.org/10.1145/2623330.2623705>

## APPENDIX A INFORMATION GAIN OF CONTEXTUAL FEATURES

	Contextuals	Defaulters	Overall
A1	0.4839	0.6444	0.5717
A2	0.4558	0.6395	0.5605
Permission	0.0040	0.0038	0.0050
Time	0.0487	0.1391	0.0130
Visibility	0.0015	0.0007	0.0010

TABLE VI  
FEATURE IMPORTANCE OF CONTEXTUAL FEATURES

## APPENDIX B INFORMATION GAIN OF BEHAVIORAL FEATURES

Feature	Importance
Amount of time spent on audio calls	0.327647825
Frequency of audio calls	0.321291184
Proportion of times screen was timed out instead of pressing the lock button	0.317631096
Number of times PIN was used to unlock the screen.	0.305287288
Number of screen unlock attempts	0.299564131
Amount of time spent unlocking the screen	0.29930659
Proportion of time spent on loud mode	0.163166296
Proportion of time spent on silent mode	0.138469725
Number of times a website is loaded to the Chrome browser	0.094996437
Out of all visited websites, the proportion of HTTPS-secured websites.	0.071096898
Number of times Password was used to unlock the screen	0.067999523
Proportion of websites requested location through Chrome	0.028404167
Time	0.019799623
The number of downloads through Chrome	0.014619351
Permission	0.001461635
Visibility	0.000162166

TABLE VII  
FEATURE IMPORTANCE OF BEHAVIORAL FEATURES