User Modeling and User-Adapted Interaction
https://doi.org/10.1007/s11257-019-09246-3

A recommendation approach for user privacy preferences in the fitness domain

Odnan Ref Sanchez[1] · Ilaria Torre[1] · Yangyang He[2] · Bart P. Knijnenburg[2]

Received: 23 August 2018 / Accepted in revised form: 7 September 2019

Abstract
Fitness trackers are undoubtedly gaining in popularity. As fitness-related data are persistently captured, stored, and processed by these devices, the need to ensure users' privacy is becoming increasingly urgent. In this paper, we apply a data-driven approach to the development of privacy-setting recommendations for fitness devices. We first present a fitness data privacy model that we defined to represent users' privacy preferences in a way that is unambiguous, compliant with the European Union's General Data Protection Regulation (GDPR), and able to represent both the user and the third party preferences. Our crowdsourced dataset is collected using current scenarios in the fitness domain and used to identify privacy profiles by applying machine learning techniques. We then examine different personal tracking data and user traits which can potentially drive the recommendation of privacy profiles to the users. Finally, a set of privacy-setting recommendation strategies with different guidance styles are designed based on the resulting profiles. Interestingly, our results show several semantic relationships among users' traits, characteristics, and attitudes that are useful in providing privacy recommendations. Even though several works exist on privacy preference modeling, this paper makes a contribution in modeling privacy preferences for data sharing and processing in the IoT and fitness domain, with specific attention to GDPR compliance. Moreover, the identification of well-identified clusters of preferences and predictors of such clusters is a relevant contribution for user profiling and for the design of interactive recommendation strategies that aim to balance users' control over their privacy permissions and the simplicity of setting these permissions.

Keywords Privacy preferences · Fitness trackers · Profiling · Privacy-setting recommendations · Privacy management · Wearable IoT devices

✉ Ilaria Torre
ilaria.torre@unige.it

Extended author information available on the last page of the article

Published online: 12 October 2019

1 Introduction

Preserving the privacy of users in the context of the Internet of things (IoT) is a growing concern. In particular, this is due to the increasing number of third party (TP) applications and personal IoT devices, and the increase in data sharing among TPs, which make privacy management more complex in the IoT. These developments not only increase privacy concerns but also make setting one's privacy preferences an increasingly complex task.

Since the field of IoT is very broad, this study focuses on personal tracking devices, which are a prominent focus in privacy research because they collect vast amounts of personal tracking data. Specifically, we will study the case of fitness trackers (e.g., Fitbit) as they are known to employ around-the-clock monitoring of users' activity.

While several frameworks and applications for managing privacy preferences have been proposed recently (e.g., Lee and Kobsa 2017; Tsai et al. 2017), most of them do not specifically concern IoT tracking data and data sharing among TPs. Building on privacy management studies in the field of ubiquitous computing, this paper aims to fill this gap by *modeling* users' privacy preferences and *recommending* privacy settings in a fitness IoT scenario.

The features for privacy preference modeling considered in this paper are based on the Privacy Preference for IoT ontology (PPIoT) (Sanchez et al. 2019) with the aim to unambiguously identify privacy preferences and data sharing permissions for both the user and TPs. The vocabulary is based on well-established ontologies for the description of privacy preferences and IoT resources; moreover, it takes into account the newly adopted EU General Data Protection Regulation (GDPR) (The European Parliament and the Council of the European Union 2016), which represents the most important change in data privacy laws in the last twenty years. While the GDPR is a significant stride toward user empowerment and control over their personal data, it requires users to make explicit decisions for every individual privacy setting. In the IoT scenario, the effort required for such explicit control can be exhaustive, especially when considering the number of devices, applications, and data collection practices that must be given individual consent by the user. Hence, our approach aims to increase its ease of use by combining the GDPR principles with the concept of *privacy recommendation*.

For the prediction and recommendation of user privacy preferences, we adopt a well-established two-step approach in user modeling. First, we aim to identify user profiles that can represent the vast diversity of privacy preferences through the use of machine learning clustering algorithms. Then, we investigate how to exploit users' privacy preferences on tracking data and personal traits to drive the recommendation of which profile best describes each user by using a tree-based classifier. The main recommendation-driving factors that we took into account include users' privacy behavior and attitudes, the negotiability of their preferences (cf. Tsai et al. 2017), as well as social influence and sociability (Wu et al. 2017), users' privacy preference feedback (cf. Tsai et al. 2017), and users' attributes (cf. "side information" Rafailidis and Nanopoulos 2016) such as demographics.

The input for our privacy preference modeling efforts comes from a crowdsourcing study on the Amazon Mechanical Turk platform. To collect a sample of fitness IoT permission settings, we simulate a fitness app prototype built to provide a semi-realistic environment, and a subsequent questionnaire. In machine learning and IoT, crowdsourcing has been successful at gathering input from a diverse set of users. While it does not always provide reliable output, we employed several selection criteria and attention check questions to ensure that we collected high-quality data from a sample that represents the fitness tracking population as accurately as possible.

Our results indicate that users' privacy attitudes (in particular privacy concerns and trust in the third party), their social behavior (in particular sociability), and the negotiability of permissions for tracking data when varying risk level or benefit (mainly sleep tracking data and phone permissions) can drive the recommendation of privacy preferences. We show that there are semantically relevant relationships between these drivers of the recommendations and the modeled privacy preference categories. However, we also demonstrate that direct privacy profile item questions (i.e., users' willingness to share their minutes of activities, provide their first name, allow access to their photographs, and share their tracking data for social purposes) provide even better predictors of their preferences.

Moreover, we show the applicability of these findings by developing recommendation strategies that simplify the task of privacy permission setting, with different levels and types of user intervention. The current application environment for our findings is our personal data manager (PDM), a framework designed to support users in managing and controlling privacy preferences with respect to third parties (Torre et al. 2016c, 2018). In the paper, we will provide details about the integration of privacy recommendations within the PDM.

The main contribution of the proposed approach is to harness users' personal tracking data permission preferences and other user traits to build user profiles and predictors of such profiles in the fitness domain. The aim is to suggest privacy settings for tracking data that fit the user's preferences, are GDPR compliant, and reduce the effort for users to set such permissions, thereby maintaining meaningful control over their preferences without unnecessary burden.

Our work applies directly to *personalized fitness services*. The issue of balancing simplicity of privacy preference setting with control of personal data processing is crucial in this domain. Based on our previous work on personal data management (Sanchez et al. 2019; Torre et al. 2016c), privacy risks (Carmagnola et al. 2014; Torre et al. 2018), and privacy profiling (Bahirat et al. 2018; He et al. 2019), the current work is novel in the following aspects: It applies the privacy profiling approach to fitness IoT (an unexplored application area), it applies the privacy profiling approach to settings data (previous work used the approach on responses to scenarios), and it is also the first work to subsequently predict cluster assignments using privacy-related questions and also indirect questions. Our findings about the predictors of privacy preferences in the specific field of fitness trackers can be used to support privacy-aware user modeling. The data model we defined for the fitness domain and the related dataset is based on popular fitness trackers; as such, it has a wide coverage of tracking data that likely includes those used by most of the personalized fitness services.

However, it is worth noting that beyond our contribution to the domain of personalized fitness services, we describe a generic method to develop user profiles and a series of recommendation strategies for privacy management. While the PPIoT vocabulary is specifically targeted to IoT, it can be substituted for other privacy management ontologies. Moreover, while we demonstrate our approach in the context of a fitness tracker, it is designed to be generalizable to other data-intensive user-centric applications.

The remainder of this paper is structured as follows. The next section presents the background and related work of this study. Section 3 discusses our research methodology together with our PDM framework. We derive our data model for the fitness domain in Sect. 4. Our method for data collection is presented in Sect. 5. The privacy profile models and their respective drivers are discussed in Sects. 6 and 7, respectively. Our recommendation strategies are presented in Sect. 8. Finally, the limitations and future work, and the concluding remarks are discussed in Sects. 9 and 10, respectively.

2 Background and related work

In this section, we first analyze the state-of-the-art literature on privacy management in mobile and IoT frameworks, and then we describe privacy preference modeling and the recommendation approaches. Finally, we discuss the representation of privacy preferences in light of the currently enforced GDPR.

2.1 Privacy management

Mobile privacy permission systems have been well studied in the literature. However, they do not properly cover the scope of new IoT devices, such as fitness trackers, that expand and extend the services and personal data that must be managed. In this section, we first provide background information about privacy permission management in mobile systems, since they serve as groundwork for the IoT context. Then, we discuss related work on frameworks for privacy management in the IoT.

2.1.1 Permission management in mobile systems

Studies in mobile privacy (e.g., Felt et al. 2012) have demonstrated that the mobile interfaces of both Android and iOS lack the potential to provide the necessary user privacy information and control (Lin et al. 2014). Several solutions have been proposed to improve mobile privacy protection and offer users more privacy control (e.g., Beresford et al. 2011). Some of these suggestions have since been taken into account to improve privacy management of current mobile systems (i.e., starting from Android 6.0+ and iOS 5.0+).

The Android permission systems can be mainly categorized as the Ask On Install (AOI) and Ask On First Use (AOFU) privacy models (Tsai et al. 2017; Wijesekera et al. 2017). In AOI[1] (Android 5.9 and below), the permissions are asked in bulk before installing a TP app. The user's option is only to allow or deny all, which affords less privacy control. Also, research shows that few users read and pay attention to the install time permissions, and even fewer understand their meaning (Felt et al. 2012; Kelley et al. 2012). These issues made room for TP apps that manage app privacy, such as Turtleguard Tsai et al. (2017) and Mockdroid Beresford et al. (2011).

[1] https://support.google.com/googleplay/answer/6014972?co=GENIE.Platform%3DAndroid&hl=en.

On the other hand, the AOFU model (Tsai et al. 2017) (Android 6.0 and above) asks permissions the first time an app uses a specific feature that needs the respective permission. In this case, users grant the permission during the actual provision of the service and will be able to weigh their willingness to share against the utility of the app. Users can also revisit and review permissions in their phone privacy settings for each app. This model makes users more informed and gives them more control (Fu et al. 2014). Moreover, it has been shown that interactive notifications are more efficient in informing users about access requests (Fu et al. 2014). The distinction of these two models is worth noting since as of July 2018, 34% of the Android users were still using the AOI model[2].

In terms of privacy management, iOS has used the AOFU model for location permission since iOS version 5.000[3], with a more comprehensive rollout in iOS 6.0 and onwards (Almuhimedi et al. 2015). Although iOS is not open source like Android, this has not stopped researchers from finding ways to improve its privacy-setting mechanism. For example, *ProtectMyPrivacy* is an app specifically designed for jailbroken iOS devices that preserves user privacy by substituting anonymized data for user data (Agarwal and Hall 2013). Although jailbreaking is deemed legal, it is not advisable to do so as jailbroken iOS devices can be used to install pirated apps that might contain privacy risks. Privacy managers specifically built for non-jailbroken iOS devices also exist, but they have reduced functionality. For example, *PiOS* (Egele et al. 2011) is a privacy manager which only has the function to check if the installed iOS apps have committed privacy breaches. Similarly, *Data Privacy Pro*[4] can only act on users' private photographs, videos, and notes.

2.1.2 Frameworks for privacy management

Several existing frameworks involve a personal data manager for privacy management. For instance, ipShield (Chakraborty et al. 2014) is a context-aware privacy framework for mobile systems that provides users with great control over their data and inference risks. My Data Store (Vescovi et al. 2015) offers a set of tools to manage, control, and exploit personal data by enhancing an individual's awareness regarding the value of their data. Similarly, Databox (Chaudhry et al. 2015) enables individuals to coordinate the collection of their personal data and make those data available for specific purposes. However, these data managers do not include user privacy profiling and recommendation in the complex IoT environment.

khealth is an IoT framework based on a personalized digital health care information system that protects users from TP adversaries (Sharma et al. 2018). Privacy can also be protected by providing different anonymity levels of data that are given to the TPs. However, it may not be possible to implement the most effective privacy standards such as data obfuscation due to numerous trade-offs and restrictions, especially in the health care and fitness domain.

[2] https://developer.android.com/about/dashboards/index.html.

[3] https://developer.apple.com/library/content/releasenotes/General/WhatsNewIniOS/Articles/iOS6.html#//apple_ref/doc/uid/TP40011812-SW1.

[4] https://itunes.apple.com/us/app/data-privacy-manager-pro-security-suit-to-lock-my-private/id625761168?mt=8

2.2 Privacy preference modeling

The study of privacy preferences is a challenging task, given the diversity of users' preferences, context conditions, and regulations. Kobsa (2001) suggested that privacy settings should be *dynamically tailored* to both legislative rules and individual user needs, since different factors affect user preferences. This is also the core principle of our approach in this paper.

In ubiquitous computing, the issue of privacy management has been studied since the early 1990s (Bellotti and Sellen 1993). In the 2000s, Brar and Kay (2004), Kay and Kummerfeld (2006), and Kay et al. (2002) focused their research on supporting user scrutiny and control over the information held by applications. In this light, *Personis* (Kay and Kummerfeld 2006; Kay et al. 2002) is a user modeling framework that ensures the user can maintain control at different levels (e.g., source identity, source type, the processes used to gather the user data, the way such information will be used to provide personalized services). Based on the same principle, Secure Personal Exchange (SPE) (Brar and Kay 2004) is a framework for personalized services and an example of privacy modeling and management in ubiquitous computing. It implements machine-processable policies based on the P3P[5] vocabulary to provide tools for representing and storing user preferences as subsets of user models (*personas*), each intended for use by particular applications. Even though the P3P became obsolete due to a lack of adoption, most of its main concepts are still being used for data protection regulations.

Context-aware privacy modeling has shown to enhance the accuracy of users' privacy preference prediction (Lee and Kobsa 2017; Wijesekera et al. 2017). Context is defined as the situation (e.g., what, when, who, where, how, etc.) under which a TP application requests access to data. The context improves the prediction; for example, when and under what circumstances the data are collected plays a big role in predicting user preferences (Wijesekera et al. 2017). Likewise, Lee and Kobsa found that the identity of the information requester (the *who* context) is an important determinant of people's privacy decisions (Lee and Kobsa 2017). Our results confirm this finding.

Leveraging the dataset collected by Lee and Kobsa, Bahirat et al. (2018) applied a data-driven design methodology to develop a privacy-setting interface and a set of smart default profiles for Internet of things devices. In the current paper, we use a similar approach to identify privacy profiles and smart default interfaces. However, while the goal of Bahirat et al. was to let users pick profiles manually, our current goal is to further classify users with respect to the profiles in order to give them a personalized profile recommendation. Moreover, our current work uses users' permissions (collected through our FitPro prototype app, based on the PPIoT and the GDPR) as a basis for privacy profiles, whereas Bahirat et al. used scenario-based input. Our current work is thus closer to a real-world implementation.

Preference modeling was also explored to enhance privacy in social networks. For instance, Facebook users are found to have 6 types of privacy profiles: privacy maximizers, selective sharers, privacy balancers, self-censors, time savers/consumers, and privacy minimalists (Knijnenburg 2017; Wisniewski et al. 2014). Moreover, in Wu

[5] Platform for Privacy Preferences https://www.w3.org/P3P/.

et al. (2016), the inclusion of both the influence of the user's social surroundings (i.e., social influence) and the future association and bond with individuals that have similar preferences (i.e., homophily effect) enhances the modeling of user preferences. These factors are also included in our study.

In the health/fitness domain, emerging sensors and mobile applications allow people to easily capture fine-grained personal data related to the long-term fitness goals. Focusing on tracker data (i.e., weight, activity, and sitting), Brar and Kay (2004) discovered that users' preferences vary by sensor [i.e., weight being the most important (Brar and Kay 2004)]. Also, their study concludes that users want to have control over their fitness data, and that they would like to have a personal copy of their data.

Modeling location privacy preference has received much attention in the literature, given the sensitivity of this information (see for instance Almuhimedi et al. 2015; Assad et al. 2007; Vicente et al. 2011; Xie et al. 2014). Assad et al. (2007) study users' preferences regarding the release of location information and provide support to differentiate their release. Vicente et al. (2011) not only consider location privacy but also absence and colocation privacy. Moreover, Xie et al. (2014) study location sharing privacy preferences with respect to different contextual parameters, including check-in time, companion, and emotion. These studies confirm that users want to control the privacy of information and that this is specifically important in ubiquitous environments.

Finally, it is worth mentioning a number of privacy preference modeling frameworks that use the semantic Web. PPO (Sacco and Breslin 2012) has pioneered modeling user's privacy preferences, giving users' fine-grained control of their preference. A survey of privacy management approaches using ontologies can be found in Perera et al. (2016). Our PPIoT ontology (Sanchez et al. 2019) that we adopt to describe our fitness data privacy model is based on PPO and other well-established ontologies for the description of preferences and the representation of IoT resources. Details on related ontologies will be provided in Sect. 4.1.

2.3 Recommendation in privacy management

Enhancing permission settings gives more control to the user, but it also increases complexity. As the number of applications that the users utilize increases (currently averaging 35 apps/user Google/Ipsos 2016), the number of permissions per application increases (currently averaging 5 permissions per app[6]), and even the number of devices users own increases (currently averaging 4 devices per user[7]), these permission models will not be enough. Indeed, research shows that burdening the user with the formidable task of setting each individual permission easily becomes a tedious task that is prone to errors (Acquisti et al. 2015; Lee and Kobsa 2017; Madejski et al. 2012). Generally, users are increasingly unable to make decisions about privacy settings due to limits in their available time, motivation, and their cognitive decision-making abilities. Moreover, users' stated privacy preferences are often inconsistent with their actual behaviors and users are likely to be uncertain about their own privacy

[6] http://www.pewinternet.org/2015/11/10/apps-permissions-in-the-google-play-store/.

[7] https://blog.globalwebindex.com/chart-of-the-day/digital-consumers-own-3-64-connected-devices/.

preferences (Lee and Kobsa 2017; Acquisti et al. 2015). In this section, we describe some of the approaches that have been proposed to solve this problem.

Privacy nudging is an effective method to increase user awareness (Almuhimedi et al. 2015). Nudging allows users to be informed about both their privacy settings and how TP applications access their data (Liu et al. 2016). Note, though, that privacy nudges lack personalization and provide only general recommendation.

Another approach that is more user-centric is *user-tailored privacy* (Knijnenburg 2017). It models users' privacy concerns and provides users with adaptive privacy decision support. This model can be seen as personalized "smart nudges" where the recommendation is aligned with the user's privacy preference. User-tailored privacy aids users in making privacy decisions by providing them privacy-related information specifically tailored to them and useful privacy control that does not overwhelm them. However, in practice it is hard to implement a general privacy model—the idea is too broad and abstract, especially given the diversity of privacy perceptions among users.

The approaches in Lin et al. (2014) and Liu et al. (2016) are closely related to our approach, but they are limited to mobile systems. The solution in Lin et al. (2014) is to provide a set of predefined privacy preference configurations. This can be attained by using machine learning algorithms to predict the best-suited preference settings for the user. It shows that within the substantial between-user variability of permission settings there exist some profiles that can collectively describe these diverse settings with substantial accuracy. These privacy profiles (Lin et al. 2014; Liu et al. 2016, 2014b) are collections of related privacy and sharing rules that correspond to privacy preferences of similar-minded users (cf. Knijnenburg 2014; Knijnenburg et al. 2013; Wisniewski et al. 2014; Xie et al. 2014). By identifying the privacy profile that matches a new user, one can provide decision support by means of a privacy recommendation (Liu et al. 2016). In Liu et al. (2014b), six privacy profiles were identified based on the analysis of 4.8 million users' privacy settings. In a subsequent paper, the authors add new features (such as the purpose of information and app categories) for modeling user privacy profiles, as well as privacy nudges that make users more aware of unexpected data practices from TPs (Liu et al. 2016).

Our two-step approach combines the profile approach and the recommendation approach and is aimed to maintain a balance between simplicity of setting and personalization, which allows users to be informed about the profile that is best suited for them and be recommended settings that are associated with their privacy preferences. Moreover, in the current paper privacy profiling is studied with specific regard to the fitness domain, a still unexplored application area.

2.4 General data protection regulation

As of May 25, 2018, the European Union (EU) enforces the General Data Protection Regulation (GDPR) (The European Parliament and the Council of the European Union 2016) which applies to the storage, processing, and use of subjects' personal data. The GDPR applies to all TPs that operate in the EU market or access data of EU residents, even if they themselves are not established within the EU. The GDPR requires users to provide explicit consent to privacy options expressed by TPs. This results in a complex task for the users, given the number of devices and applications they interact with, each of which will have a consent procedure which will have to be processed specifically.

Our PDM uses the vocabulary of the PPIoT ontology (Sanchez et al. 2019) which includes concepts and properties that address the GDPR requirements for the management of personal data, including the reason, method, and persistence of data access, and the maximum retention period of data in the hand of the accessing parties. Besides PPIoT, other ontologies address GDPR concepts for different purposes. The ontologies proposed in Palmirani et al. (2018), Pandit et al. (2017, 2018) address the formal representation of GDPR articles, TP obligations, and provenance modeling, respectively. Another ontology was proposed in Elluri et al. (2018) to represent GDPR rules concerning cloud data and addressing the obligations of consumers and providers. Compared to the mentioned ontologies, the PPIoT ontology captures IoT privacy permissions more comprehensively and is focused on user privacy preferences.

3 Research methodology

3.1 Personal data manager

As stated in Sects. 1 and 2, previous research on privacy needs and the new GDPR requirements call for an increase in users' control over the storage, processing, and sharing of personal data. Such control is very complex for IoT scenarios, though, which is why the privacy-setting task must be simplified to accommodate the limited cognitive abilities of the users and the risk of errors, as discussed in the related works. Our proposals for supporting users in such a task, while general, are deployed in the application environment of our personal data manager (PDM) framework described in detail in various previous works (cf. Torre et al. 2016a, b, c, 2018). PDM is designed to be an intermediary between the user, her/his devices through their dedicated third parties (TPs), and other third parties (i.e., fourth parties) that want access to her/his data (Fig. 1). The PDM is mainly responsible for managing the interaction among these entities, the access control for TPs (authentication, authorization, privacy policy evaluation), and the user's privacy preference settings.

3.2 Research questions

In the intended environment discussed above, the main issue we aim to address is the following: How can we effectively support users in setting their privacy permissions. The settings need to be sufficiently granular to be GDPR compliant and to fit users' real preferences. But they should also be simple enough to allow users to maintain meaningful control over their settings. Our approach is to provide users semiautomated interactive *privacy recommendations*. In this paper, we investigate how the PDM can recommend personalized privacy profiles in a simple, usable way. To attain this goal, we formulate four research questions for privacy preference recommendation.

RQ1 How can we formally represent users' privacy preferences in a way that is suitable for both the user and the TP?

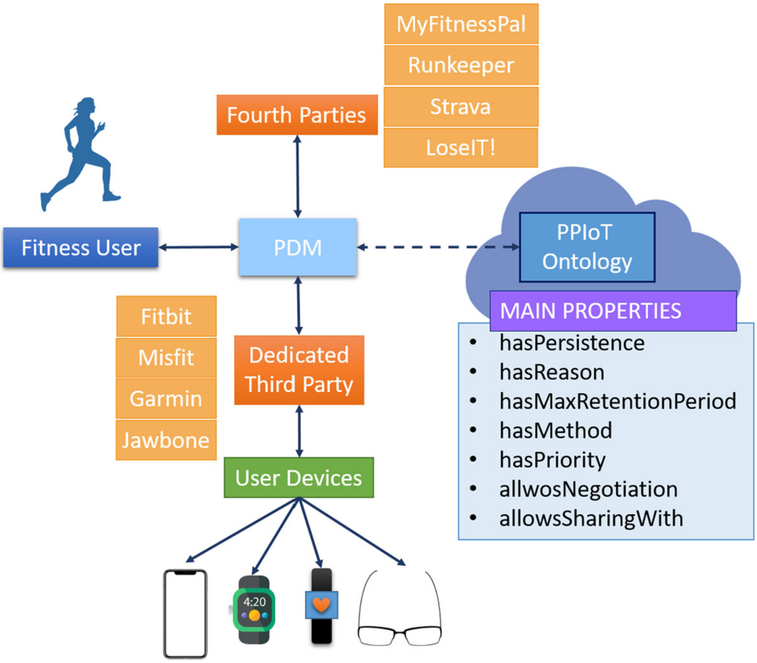


Fig. 1 Personal data manager (PDM) framework

To answer this question, we propose a fitness data privacy model which captures the general data permissions used by fitness trackers and adopts our PPIoT vocabulary to handle disputes on different and conflicting privacy terminologies.

RQ2 Is it possible to identify well-defined privacy profiles that can represent the diversity of users' privacy preferences?

To answer this question, we conduct an unsupervised machine learning analysis (clustering) to cluster users' privacy settings into distinct profiles by recruiting a total of 310 Fitbit users. We collect privacy profile data by developing a fitness app installation simulator (FitPro) that captures the user privacy permission settings, followed by a questionnaire. Our dataset is collected through the Amazon Mechanical Turk platform.

RQ3 Are there any privacy profile items or questionnaire items that can be used to predict which privacy profile best describes a user?

To answer this question, we conduct a supervised machine learning analysis (tree learning) to find privacy profile items and questionnaire items (i.e., privacy attitude, negotiability, social behavior, exercise tendencies, and demographics) that best predict the user profiles from RQ2. Users' answers to these settings/questions subsequently allow us to provide them an accurate recommendation of the profile most suitable as a starting point for their privacy settings.

RQ4 How can we effectively exploit the results to provide privacy profile recommendations?

To answer this question, we develop a series of recommendation strategies and related user interfaces based on the machine learning results. We aim to integrate

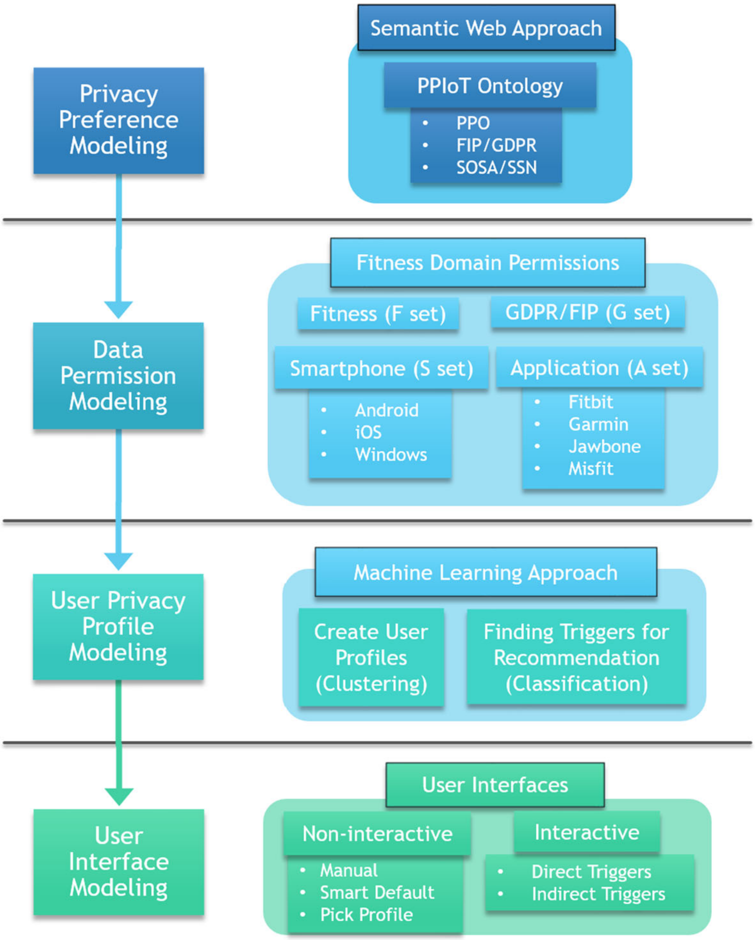


Fig. 2 Inter-model research work flow

these recommendation strategies within our PDM framework to balance control of privacy information and simplified management.

The workflow of this study is shown in Fig. 2. First, we describe the privacy preference modeling using a semantic Web approach and develop a privacy model for fitness data in the IoT domain, which answers **RQ1**. Then, we create user profiles and find the determiners for prediction, which answers **RQ2** and **RQ3**, respectively. Finally, we develop recommendation strategies and the related user interfaces based on these results, which answers **RQ4**. The layer stack in the figure is based on sequence: Each layer defines a high-level conceptual model, and the unidirectional links define their sequential relationships.

4 Fitness data privacy model

Fitness trackers collect around-the-clock user activity, which puts them among the IoT devices that capture the most intimate information about their users. This section outlines which permissions are asked by a variety of fitness trackers. We then derive our data model for the IoT fitness domain from this analysis. To address RQ1, we exploit the *vocabulary of the PPIoT ontology* to unambiguously represent users' privacy preferences, including GDPR-based preferences.

4.1 PPIoT vocabulary

The Privacy Preference for IoT (PPIoT) ontology (available online[8]) (Sanchez et al. 2019) has been designed to fill a gap in privacy preference description for the IoT, by combining and extending existing ontologies for privacy preference and for IoT. Specifically, PPIoT is based on the Privacy Preference Ontology (PPO[9]) and the W3C Semantic Sensor Network Ontology (SOSA/SSN[10]); moreover, it addresses the GDPR requirements (see Sect. 2.4) and incorporates the fair information practices (FIP) principles and state-of-the-art recommendations for privacy protection in the IoT data sharing context.

Among the concepts in the PPIoT vocabulary, in this study we refer mainly to those described below.

(i) Personal data—*Dataset* in the PPIoT vocabulary, is mapped to the personal data concept in Art. 4 of the GDPR. It represents the personal data for which privacy permissions can be expressed. The concept is general enough to be instantiated in different domains.

(ii) The owner of the personal data—*User* in the PPIoT, represents the data subject in Art. 4 of the GDPR.

(iii) The TP that requests to access/process the personal data—*Entity* in the PPIoT, addresses the third party, controller, processor, and recipient in Art. 4 of the GDPR[11].

(iv) The privacy preference conditions of the user for giving consent to TPs to access/process her data —*Condition* in the PPIoT, allows the specification of conditions for consent using the *properties* of the condition concept. Figure 1 shows the main condition properties from the PPIoT vocabulary: hasMaxRetentionPeriod, allowSharingWith, etc. Note that the object of these properties is modeled with further PPIoT concepts, such as the type of entity that requests access—*EntityType* in the PPIoT.

In our framework, the vocabulary of the PPIoT ontology is used by the PDM for representing the user's privacy preferences. Below, we will provide details for its use in the fitness data privacy model.

[8] http://pdm-aids.dibris.unige.it/PPIoT.

[9] http://vocab.deri.ie/ppo.

[10] https://www.w3.org/ns/ssn/.

[11] The distinction among such subjects in the GDPR, which clarifies the legal obligations of the TP, is not relevant to the aim of a user-side privacy manager.

4.2 Datasets of the fitness data privacy model

Table 1 shows a comparison of the data collection practices of four fitness trackers. We selected these trackers based on both popularity and maturity of their software solutions (i.e., we only selected those that have a working Web API enabling them to share/integrate with third parties). A complete list of popular fitness trackers that have resource APIs can be found in Zhao et al. (2016). Among these fitness trackers, in this study we considered Fitbit, Garmin, Jawbone, and Misfit.

As shown, the requested data can be categorized into three sets. The first set of requested data are the smartphone permissions, which are requested during installation or the first use of the app. We define this set as the *S set*. The next set of data is requested inside the fitness tracker application, usually as the user signs up for the app's online services. We define this set as the *A set*. Finally, the app collects fitness data during the use of the tracker, which we define as the *F set*. Importantly, the data items in the F set are by default only available in the tracker's own application, but other TPs can ask for permission to gain access to this data.

The final column in Table 1 is the superset of requested data items collected by the four trackers, taking into account the different mobile operating systems. Moreover, it includes the *G set* which concerns the GDPR requirements. It will be explained in Sect. 4.6. The data items in this final column form the *Fitness Data Privacy Model* for this study. The permissions are requested in the order S-A-F-G, which will be used throughout this paper.

In the next sections, we describe the fitness datasets for the privacy model. The data items in the S-A-F sets are instances of the PPIoT *Dataset* concept introduced in the previous section, while the data items in the G set are the objects of the privacy condition properties mentioned in the previous section and further detailed in Sect. 4.6.

4.3 The S set (smartphone permissions)

The request of smartphone permissions differs not only by fitness tracker but also by mobile OS. We took into account Android, iOS, and Windows Mobile, acknowledging that Android permissions changed from "ask on installation" (AOI) in version 5.9 and below to "ask on first use" (AOFU) in version 6.0 and above. While Table 1 considers the Android AOI permissions requested by various fitness apps, Fig. 3a, b shows the Fitbit's permissions for Android AOFU and iOS for comparison.

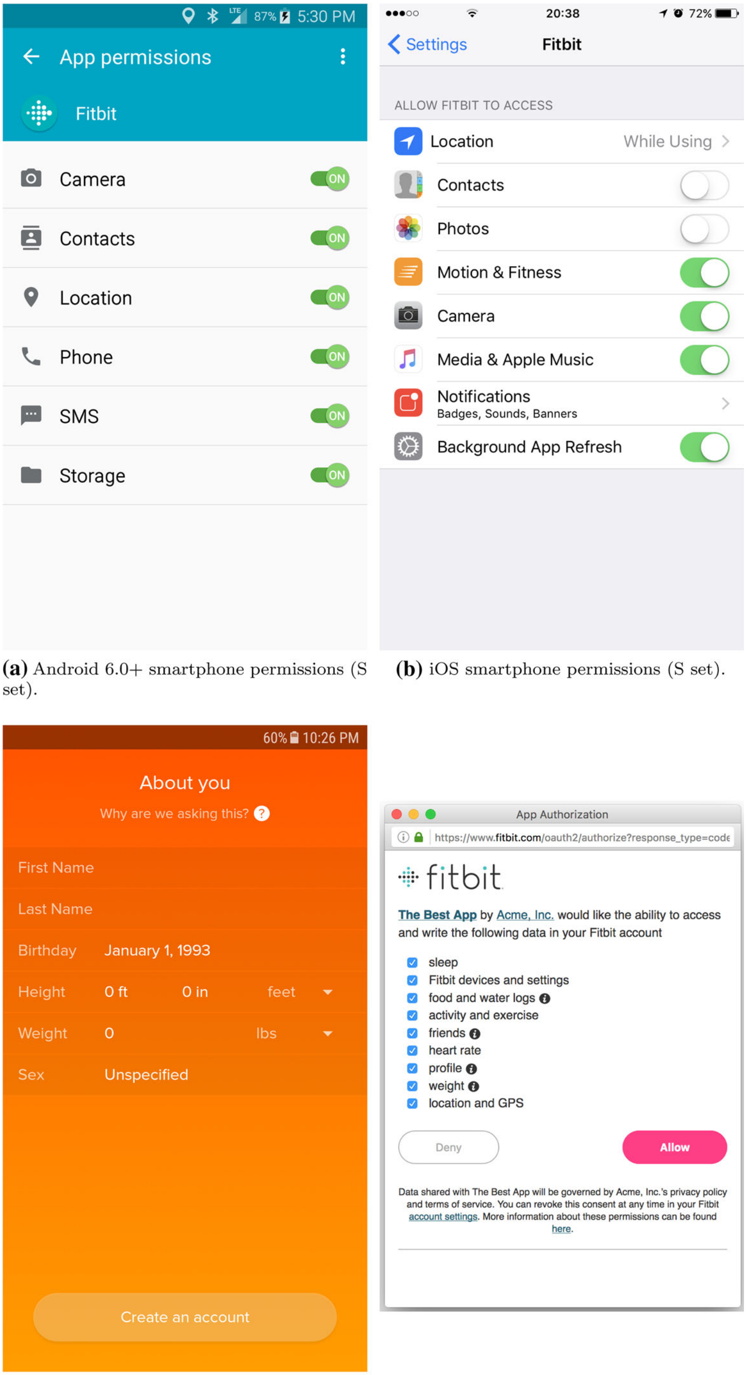
The final data model for the S set is composed of the permissions requested by the fitness apps in Table 1. The *background App* and *Notifications* iOS phone permissions are not taken into account since these permissions are not relevant for third-party data access. *Other* permissions in Android AOI are also not taken into account, except for *Bluetooth*, which older versions of Android put into this category. The *Device & Call information* is known in newer Android versions as the *Phone* permission and is included as such. The permission for *Photos/Media/Files* in Android AOI was divided into *Photos* and *Media & Music* to reflect the granularity of iOS permissions. We finally have a total of 12 permissions in the S Set.

Table 1 Comparison of permissions asked by fitness trackers and the fitness data privacy model used in this study

	Fitbit	Garmin	Jawbone	Misfit	Our data model
(S Set) Smartphone Permissions	Bluetooth	Bluetooth	Bluetooth	Bluetooth	Bluetooth
	Camera	Calendar	Camera	Camera	Camera
	Contacts	Camera	Contacts	Contacts	Contacts
	Device & Call Inf.	Contacts	Device & Call Inf.	Device & Call Inf.	Identity
	Identity	Device & Call Inf.	Identity	Identity	Location
	Location	Identity	Location	Location	Media & Music
	Photographs/Media/Files	Location	Microphone	Phone	Mobile Data
	SMS	Phone	Phone	Photographs/Media/Files	Motion & Fitness
	Storage	Photographs/Media/Files	Photographs/Media/Files	SMS	Phone
		SMS	SMS	Storage	Photographs
		Storage	Storage		SMS
		Wifi Inf.			Storage
(A set) In-app Requests	Birth date	Birth date	Birth date	Birth date	Birth date
	Gender	Gender	Gender	Gender	Gender
	Height	Height	Height	Height	Height
	Name (First)	Name (Display)	Name (First)	Name (Full)	Name (First)
	Name (Last)	Name (Full)	Name (Last)	Occupation (Optional)	Name (Last)
	Weight	Weight	Weight	Weight	Weight
(F Set) Fitness Data	Activity & Exercise	Full Fitness data	Basic Info	Device	Activity & Exercise
	Activity minutes	Location	Extended Info	Goal	Activity minutes
	Calories activity	Sync Device	Heart rate	Profile	Calories activity
	Distance		Meals	Session	Distance
	Elevation		Moves	Sleep	Elevation
	Floors		Sleep	Summary	Floors

Table 1 continued

	Fitbit	Garmin	Jawbone	Misfit	Our data model
	Steps		Friends list	Activity calories	Steps
	Devices & Settings			Calories	Devices & Settings
	Food & Water Logs			Distance	Food & Water Logs
	Friends			Steps	Friends
	Heart rate				Heart rate
	Location & GPS				Location & GPS
	Profile				Profile
	Sleep				Sleep
	Weight				Weight
					Entity Types
					Fitness/Health apps
(G Set) GDPR Permissions					Fitness programs (corp.)
					Fitness programs (gov't.)
					Social Networks (friends)
					Social Networks (public)
					Other apps
					Purposes
					Convenience
					Commercial
					Health
					Safety
					Social
					Frequency
					Retention



App permissions

Fitbit

- Camera ON
- Contacts ON
- Location ON
- Phone ON
- SMS ON
- Storage ON

(a) Android 6.0+ smartphone permissions (S set).

Fitbit

ALLOW FITBIT TO ACCESS

- Location: While Using >
- Contacts: off
- Photos: off
- Motion & Fitness: on
- Camera: on
- Media & Apple Music: on
- Notifications: Badges, Sounds, Banners >
- Background App Refresh: on

(b) iOS smartphone permissions (S set).

About you

Why are we asking this? ?

First Name

Last Name

Birthday January 1, 1993

Height 0 ft 0 in feet

Weight 0 lbs

Sex Unspecified

Create an account

(c) In-app requests (A set).

The Best App by Acme, Inc. would like the ability to access and write the following data in your Fitbit account

- ☑ sleep
- ☑ Fitbit devices and settings
- ☑ food and water logs
- ☑ activity and exercise
- ☑ friends
- ☑ heart rate
- ☑ profile
- ☑ weight
- ☑ location and GPS

Deny Allow

Data shared with The Best App will be governed by Acme, Inc.'s privacy policy and terms of service. You can revoke this consent at any time in your Fitbit account settings. More information about these permissions can be found here.

(d) Fitness data sharing permissions (F set).

Fig. 3 Examples of permission requests: Fitbit permissions

4.4 The A set (in-app requests)

In addition to the smartphone permissions that fitness tracking apps ask for, they also gather information inside their application, e.g., as part of the sign-up process for their online services. These data usually include the user's *First Name* and *Last Name*, *Birth Date*, *Gender*, *Height*, and *Weight*. Note that these data items are mandatory for all fitness trackers in Table 1; the only optional piece of information is Misfit's request for the user's *Occupation*. Figure 3c shows the *A set* for the Fitbit app (other apps are similar), as reported in Table 1. A total of 6 permissions are considered as part of the A set.

4.5 The F set (fitness data)

The F set contains the data fitness trackers collect while the user is using the device. Some of these data are automatically collected by the tracker (e.g., steps, distance) and shared with the device's own fitness tracking app (e.g., the Fitbit device shares fitness data with the Fitbit app), while the user has to enter other data manually into the app (e.g., food and water logs, friend list).

While these data are "shared" with the native fitness tracker TP by default (since this TP serves as the collecting TP), most trackers have an API that allows users to further share these data with other TPs in exchange for additional fitness or health services the user can benefit from. This data sharing was modeled in Torre et al. (2018) together with its associated risks. Table 1 shows the data that can be shared to other TPs from the four considered fitness apps. In this comparison, Fitbit gives the users more granular control over which of the fitness data can be shared with other TPs through their API, as shown in Fig. 3d[12]. Additionally, these settings can be revisited in their Web app[13], where users have the option to revoke access. The other apps in Table 1 also give users control but only give them the option to allow/deny the other TP access to the entire F set. We follow Fitbit's permission model for this set but give users even more fine-grained control over *Activity and Exercise* data, breaking these permissions down into steps, distance, elevation, floors, activity minutes, and calories activity. We implement this additional granularity because these data involve a particular inference risk, potentially exposing some of the other data in this set (Torre et al. 2018). A total of 14 permissions are included in the F set.

Note that the F set permissions are repeated for *each additional TP* that requests access to this data. As such, these permissions are not for the native app of the fitness tracker, but for other TP apps that the user desires to use and allow access to her/his fitness tracking data. In this study, instead of taking into account individual TPs, we use the PPIoT *EntityType* concept, mentioned in Sect. 4.1, to investigate which category of TP apps (i.e., "who") the user prefers to share with. This parameter has been shown to be important in determining users' privacy preferences (Lee and Kobsa 2017). Since entity types are intimately related to GDPR-based requirements, these permissions are included in the G set.

[12] https://dev.fitbit.com/build/reference/web-api/oauth2/.

[13] https://community.fitbit.com/t5/Flex-2/How-do-I-revoke-access/td-p/2701359.

4.6 G set (GDPR-based permissions)

The G set includes permissions that are based on GDPR requirements and modeled using the PPIoT vocabulary. Below, we list the PPIoT properties for expressing permissions for accessing the user data (also reported in Fig. 1). These properties will subsequently be explained with reference to the GDPR. The main permissions concern the frequency (*hasPersistence*), reason (*hasReason*), and method (*hasMethod*) of data access, as well as the retention period of collected data (*hasMaxRetentionPeriod*). Other controls provided to users to manage their privacy include the location of the device (*hasSensingLocation*), the priority of a certain preference (*hasPriority*), preferences for further sharing of tracking data (*allowsSharingWith*), negotiability of the privacy condition (*allowsNegotiation*), and types of TPs that can request access to their data (*EntityType*).

GDPR Article 3-1 and 3-2 (The European Parliament and the Council of the European Union 2016), respectively, state that the regulation:

- "applies to the processing of personal data in the context of the activities of an establishment of a controller or a processor in the Union, regardless of whether the processing takes place in the European Union (EU) or not," and;
- "applies to the processing of personal data of data subjects who are in the Union by a controller or processor not established in the Union."

Article 3-1 means that any TP registered in the EU must obey this regulation, regardless of doing the processing within the EU or not. Additionally, Article 3-2 means that any TP, whether or not registered in EU, that processes personal data of subjects who are in the EU must also abide to this regulation. Therefore, GDPR's territorial scope is not limited to the EU; it also applies to TPs outside the EU who access personal data of subjects who are in the EU. For this reason, we made sure that our PPIoT vocabulary conforms with the GDPR.

The data handling principles are defined in Art. 5 of the GDPR (The European Parliament and the Council of the European Union 2016). It specifies that TPs must declare explicit and legitimate purposes ("purpose limitation"). Consequently, our *hasReason* enforces TPs to explicitly provide the underlying purpose of their data access.

Personal data must also be kept for no longer than is necessary ("storage limitation"). It is also stated in Art. 15 that the TPs must provide the envisaged period for which the personal data will be stored, or, if not possible, the criteria that will be used to determine that period. This regulation is taken into account by the *maxRetentionPeriod* property.

GDPR also requires data handling to be adequate, relevant, and limited to what is necessary ("data minimization"). TPs' frequency of data access has thus to be adequate and limited to what is necessary; this is addressed by the (*hasPersistence*) property.

Data accessed must also be processed in a way that ensures appropriate security of the personal data, e.g., using appropriate technical or organizational measure ("integrity and confidentiality"). For this reason, the *hasMethod* property requires TPs to state the accessing methods to meet this requirement.

It is stated in Art. 15 that the accessing TPs must also specify the recipients or categories of recipients to whom the personal data have been or will be disclosed. The *EntityType* property conforms to this regulation as it provides categories of TPs that receive personal data.

We report the terms used to unambiguously represent these permissions. The purpose of data collection, *hasReason*, includes *safety*, *health*, *social*, *commercial*, and *convenience*. The frequency of data access, *hasPersistence*, includes *continuous access*, *continuous access but only when using the app*, and *separate permissions for each workout*. For the retention period of collected data, *hasMaxRetentionPeriod*, permissions include *retain until no longer necessary*, *retain until the app is uninstalled*, and *retain indefinitely*.

The types of TPs (instances of *EntityType*) that can request access to the user's Fitness data include *Fitness/Health apps*, *Social Network (SN) apps* (*public* or *friends* only), *corporate* and *government Fitness Programs*, and *other apps* on the user's phone.

We did not include the *hasMethod* property since it involves technical background, as stated in Sect. 2.4, which may not be known to the users. For simplicity, we assume that the TPs' *hasMethod* data access is *encrypted*.

4.7 A conundrum of settings

We note that Fitbit asks for a staggering total of 24 permissions across the S, A, and F data sets. Our data model, which takes a superset of permissions asked by all four fitness trackers, more granular *Activity and Exercise* data, and the additional G set, include 45 permissions in total. Moreover, if users want to share their fitness data (F set) with one or more additional health or fitness tracking apps, the permissions for this must be decided upon for each additional TP individually.

Most current fitness tracker apps do not ask these permissions in a clear way, and the settings are often hard to find in case the user wants to change them. That said, even with a more usable UI for making these settings, the sheer number of them is arguably a significant burden to the user and cause of possible errors. This is why we advocate the use of semiautomated interactive *privacy recommendations* to partially relieve users' burden of setting each of these individual permissions and meanwhile maintain the control on privacy preferences.

5 Data collection

To collect a sample of fitness IoT permission settings, we recruited 310 participants through Amazon Mechanical Turk. After data preprocessing, we utilized the data of 295 participants. We asked people to only participate if they were active Fitbit users[14], and checked this requirement by asking participants to enter the first and last few digits of their Fitbit serial number. The participants consisted of 34.2% males and 65.8%

[14] We restricted our study to Fitbit users rather than users of any fitness trackers to make sure that our sample had a more homogeneous existing experience with fitness permission setting interfaces.

females, had mean age of 35, and were generally highly educated (62% had at least a bachelor's degree). We restricted our study to fitness tracker users to detect the real preferences of target users.

We developed a prototype fitness app named *FitPro*, which systematically asked for all of the permissions in the fitness data privacy model that we defined in Sect. 4 (see Table 1, last column). Each participant used this prototype, followed by a questionnaire.

5.1 FitPro prototype fitness app

The goal of the *FitPro* prototype fitness app is to collect privacy preferences of the participants in a semi-realistic environment[15]. As shown in Fig. 4, the permission-setting interface of *FitPro* consists of the following parts (the order of the screenshots follows the prototype):

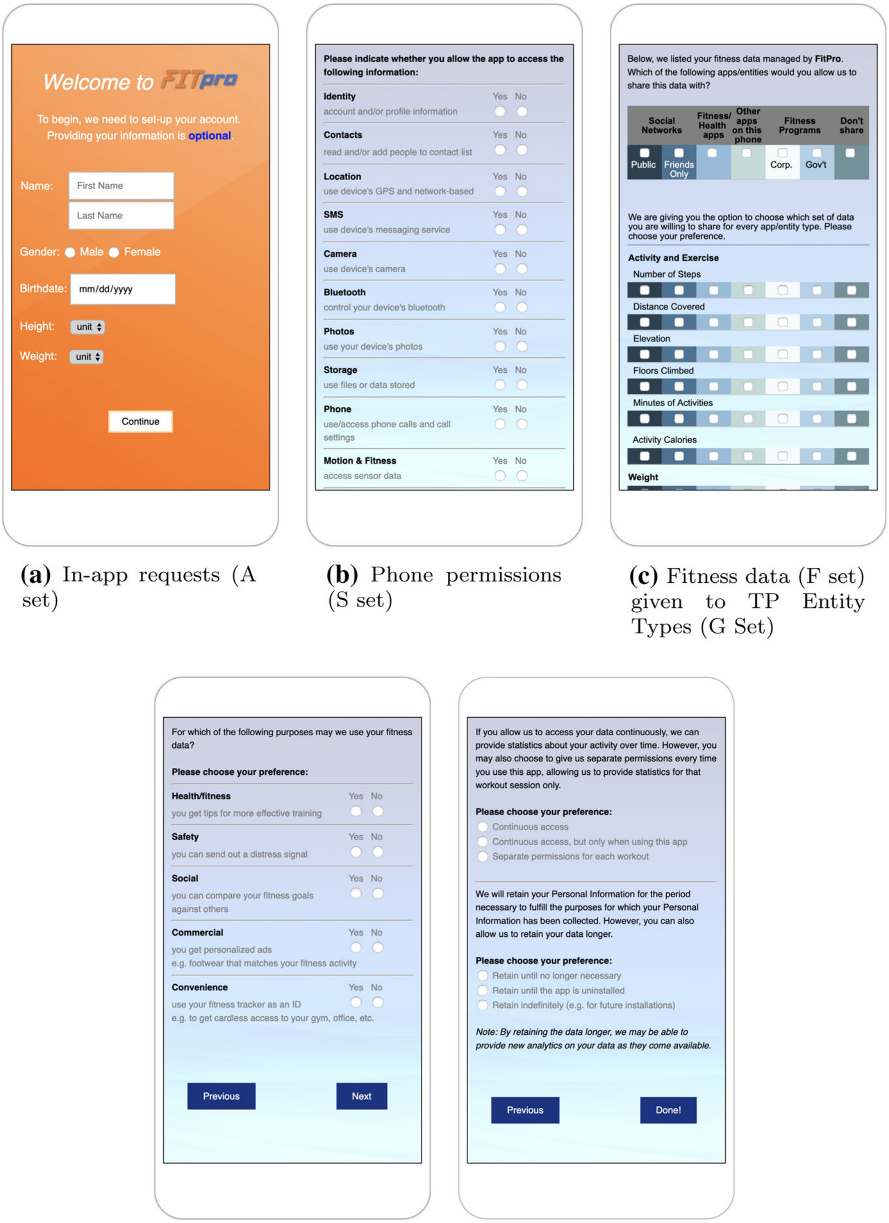
- Figure 4a shows the permissions UI for the *A set*—the data users are asked to provide as they first open the app and sign up for the fitness tracker's services. In most existing fitness trackers (including the ones we discussed in Sect. 4), these data are mandatory. In our simulation, they are optional, allowing us to measure whether participants would decide to withhold any of these data.
- Figure 4b shows the permissions UI for the *S set*—the permissions the TP needs from the smartphone. These permissions are usually asked all at once on installation or one-by-one on first use, but we decided to integrate them into our permission-setting interaction by requesting them on a separate screen in our *FitPro* app.
- Figure 4c shows the UI for the permissions to share the collected fitness data (*F set*) with other TP entity types (*G set*); as such, this UI combines the requests about "what data" can be accessed by "who". As discussed in Sect. 4.5, sharing fitness data with other apps is a common phenomenon; 40.33% of the participants in our study indicated that they had permitted other apps to access their fitness data. Rather than setting these permissions on an ad hoc basis per requesting app, our prototype allows the user to set these permissions for the various types of entities defined in Sect. 4.6.
- Figure 4c, d shows the UI screens for the *G set* permissions which address the GDPR requirements—specifically, the allowed purposes for which data may be accessed, and the frequency and retention period of the accessed data, respectively.

5.2 Questionnaire

After using the prototype, we asked participants to fill out a questionnaire[16]. The goal was to investigate if certain user traits, some of them already studied in the literature, have relations with participants' privacy behaviors collected through FitPro. Specifically, we aimed to measure participants' privacy-related attitudes (trust, privacy

[15] The prototype can be used at http://pdm-aids.dibris.unige.it/simulation.php.

[16] http://pdm-aids.dibris.unige.it/questionnaire.php.



Welcome to FITpro

To begin, we need to set-up your account. Providing your information is optional

Name: First Name

Last Name

Gender: ○ Male ○ Female

Birthdate: mm/dd/yyyy

Height: unit

Weight: unit

Continue

(a) In-app requests (A set)

Please indicate whether you allow the app to access the following information:

	Yes	No
Identity account and/or profile information	○	○
Contacts read and/or add people to contact list	○	○
Location use device's GPS and network-based	○	○
SMS use device's messaging service	○	○
Camera use device's camera	○	○
Bluetooth control your device's bluetooth	○	○
Photos use your device's photos	○	○
Storage use files or data stored	○	○
Phone use/access phone calls and call settings	○	○
Motion & Fitness access sensor data	○	○

(b) Phone permissions (S set)

Below, we listed your fitness data managed by FitPro. Which of the following apps/entities would you allow us to share this data with?

Social Networks		Fitness/ Health apps	Other apps on this phone	Fitness Programs		Don't share
□ Public	□ Friends Only	□	□	□ Corp.	□ Gov't	□

We are giving you the option to choose which set of data you are willing to share for every app/entity type. Please choose your preference.

Activity and Exercise

Number of Steps	□	□	□	□	□	□	□
Distance Covered	□	□	□	□	□	□	□
Elevation	□	□	□	□	□	□	□
Floors Climbed	□	□	□	□	□	□	□
Minutes of Activities	□	□	□	□	□	□	□
Activity Calories	□	□	□	□	□	□	□

Weight

(c) Fitness data (F set) given to TP Entity Types (G Set)

For which of the following purposes may we use your fitness data?

Please choose your preference:

	Yes	No
Health/fitness you get tips for more effective training	○	○
Safety you can send out a distress signal	○	○
Social you can compare your fitness goals against others	○	○
Commercial you get personalized ads e.g. footwear that matches your fitness activity	○	○
Convenience use your fitness tracker as an ID e.g. to get cardless access to your gym, office, etc.	○	○

Previous Next

(d) Purpose of collection (G set)

If you allow us to access your data continuously, we can provide statistics about your activity over time. However, you may also choose to give us separate permissions every time you use this app, allowing us to provide statistics for that workout session only.

Please choose your preference:

○ Continuous access
○ Continuous access, but only when using this app
○ Separate permissions for each workout

We will retain your Personal Information for the period necessary to fulfill the purposes for which your Personal Information has been collected. However, you can also allow us to retain your data longer.

Please choose your preference:

○ Retain until no longer necessary
○ Retain until the app is uninstalled
○ Retain indefinitely (e.g. for future installations)

Note: By retaining the data longer, we may be able to provide new analytics on your data as they come available.

Previous Done!

(e) Freq. & Retention (G set)

Fig. 4 Our prototype fitness app, presenting UIs for collecting fitness tracker privacy settings

concerns, perceived surveillance and intrusion, and concerns about the secondary use of personal information), the negotiability of their privacy settings, their social behavior (social influence and sociability), exercise tendencies (a proxy for their attitude and knowledge about fitness tracking), and demographic information. The questions used in this study are presented in Table 4 in the Appendix.

5.2.1 Privacy attitude

Our privacy attitude questions consist of 5 topics that were used to study different attitudes. Questions on participants' trust in app provider were derived from Knijnenburg and Kobsa (2013) and Sutanto et al. (2013). Questions on general privacy concerns are based on (Malhotra et al. 2004) which was originally based on Smith et al. (1996). Items regarding participants' perception of surveillance and intrusion, and secondary use of personal information are taken from Xu et al. (2008, 2012), and Smith et al. (1996), respectively. These user attitudes are used extensively in the privacy literature and are proven to have significant effects on users' privacy behaviors.

5.2.2 Negotiability of privacy settings

Users' preferences are rarely static, and users' "preference dynamics" (i.e., the rate at which their preferences evolve) tend to differ per person and per domain (Rafailidis and Nanopoulos 2016). Moreover, in the field of privacy, users' decisions tend to depend on the risks and benefits of disclosure (Knijnenburg et al. 2013). Following this approach, we take the negotiability of participants' privacy settings into account in this study. We measure it as an event-based change of preference: We ask participants to re-assess their disclosure decision for each item in the S, A, and F sets, imagining that the benefits or risks of disclosure increase or decrease (i.e., four re-assessments for each item).

5.2.3 Social behavior

Research shows that users' activities and preferences are to a certain extent affected by the social network around them (Wu et al. 2017). Conversely, the homophily effect suggests that people form associations with individuals that have similar preferences (Wu et al. 2017). Similarly, sociability (i.e., the ability to interact) also is a factor that can be used to predict links between users (Si et al. 2017) who do not necessarily have similar preferences. We explored this dynamic as a potential motivator for sharing one's exercise activity by creating a questionnaire regarding social influence and sociability in the fitness domain.

5.2.4 Exercise tendencies

These questions are grouped into two topics: exercise attitudes and healthy living expertise. The former items are fully self-developed. Our aim is to investigate if participants' exercise attitude (e.g., the intensity of exercise, type of exercise, their health, how important exercise is to them, the reason for exercising) influences their tendency to allow fitness apps to collect and share their data. The healthy living expertise questions are taken from Knijnenburg (2015) and measure how knowledgeable participants are about fitness tracking. Domain experts tend to be less concerned about their privacy than domain novices; hence, we expect an association between these questions and participants' privacy settings.

5.2.5 User demographics

User demographics such as gender, age, location, and education are often used to improve recommendation accuracy (Rafailidis and Nanopoulos 2016). In our questionnaire, we introduce this category to investigate if there is an association between participants' privacy settings and their demographic attributes, as resulted in previous studies (cf. Knijnenburg and Kobsa 2013).

6 Privacy preference profiles

In this section, we present our data analysis, describe our method of clustering privacy settings, and generate cluster-based privacy-setting profiles. The analysis in this section is intended to address **RQ2**.

6.1 Data analysis

Figure 5 shows that there is considerable variability in the average rate at which each permission is allowed or denied in our sample. The permissions requested by the application (A set), mainly concerning demographics (see Table 1), have a high disclosure rate, which is in line with the results of other studies (cf. Knijnenburg and Kobsa 2013).

For the smartphone permissions (S set), participants are more likely to allow motion, location, Bluetooth, and mobile data. This makes sense, because these are the minimum permissions needed to run a fitness tracker app. In this set, the permission to access photographs or contacts is granted much less often.

Regarding the purpose, frequency, and retention period of data collection (G set), participants seem open to data collection for health (the main purpose of a fitness tracker) and safety (another purpose often indicated by fitness trackers for continuous location-tracking services). Conversely, users are less likely to agree to data collection with an indefinite retention period, and they prefer not to share data with government fitness programs or publicly on social media.

We do not show the fitness data (F set) in Fig. 5 because the permissions for this data set are requested for multiple entity types of the G set, as discussed in Sect. 4.5. Hence, we present these data items in Fig. 6 instead, showing each permission for each *EntityType*. Participants are more likely to give permission to share their data with their friends on social networks and to other health/fitness apps, and they are less likely to give permission to share their data with government fitness programs or publicly on social media. As for various data types, steps are shared most openly, while location, friends, and weight are shared less openly.

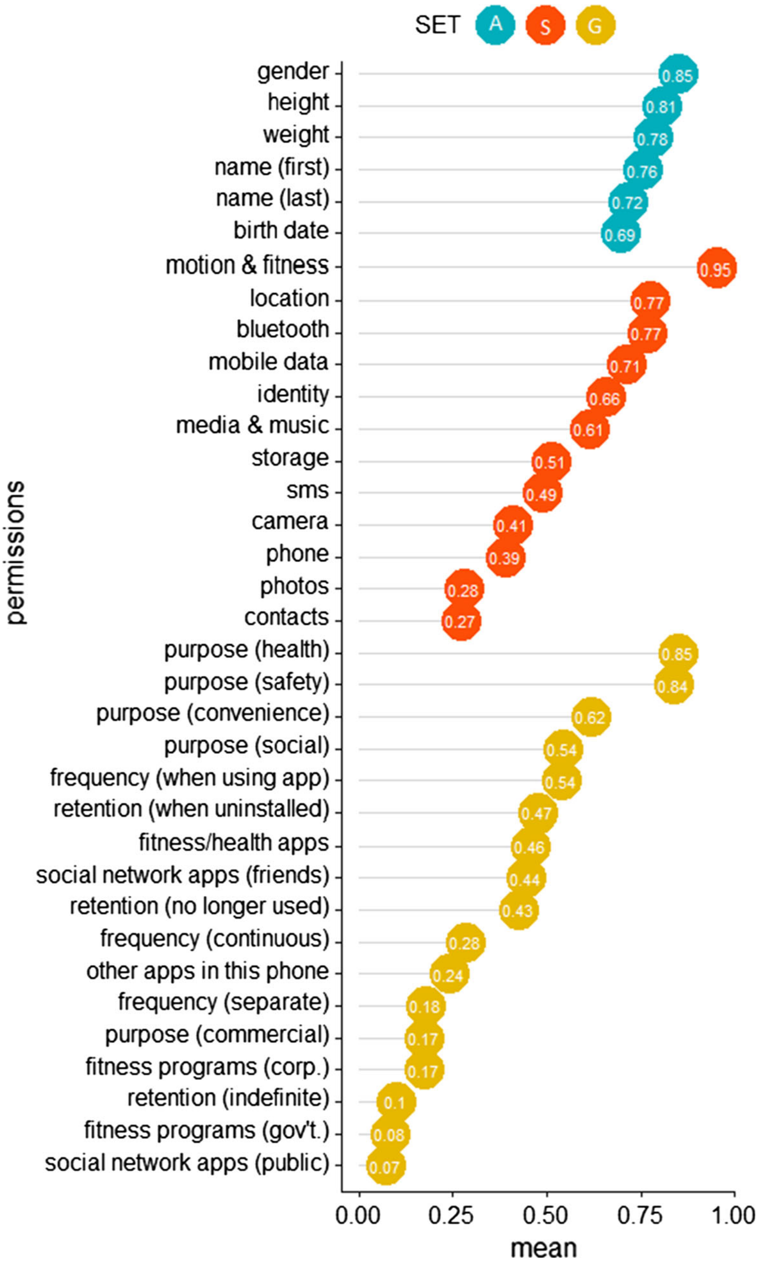


Fig. 5 Average acceptance rates of each privacy permission (allow = 1, deny = 0)

Upon further inspection, we note that participants tend to share either (almost) all or (almost) none of their fitness data with a given entity. This suggests that fitness data permissions are more likely to be influenced by the receiver ("who") rather than the

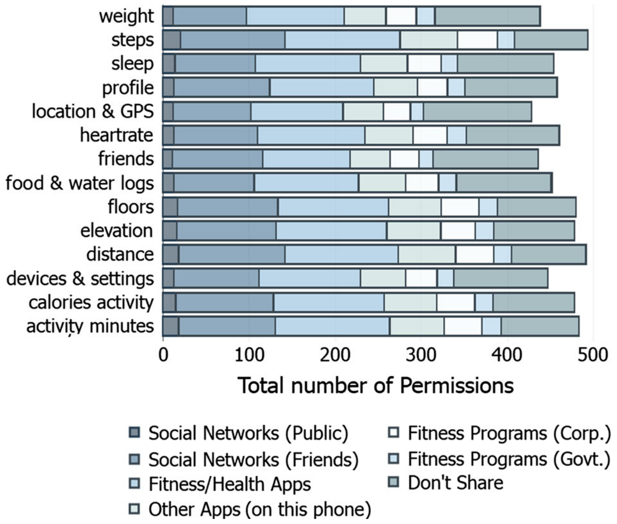
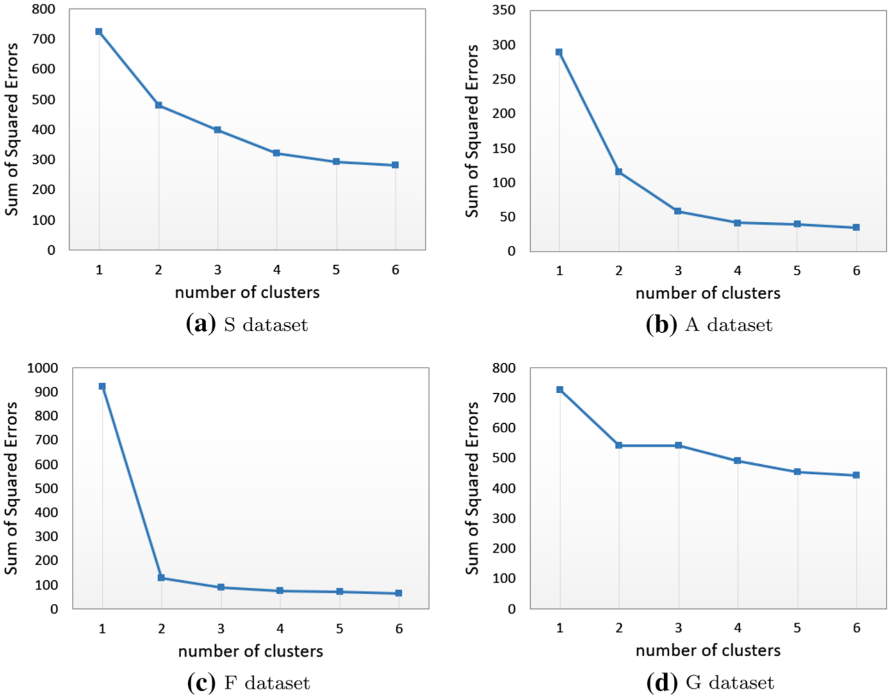


Fig. 6 Distribution of fitness data (F set) with respect to different entity types (G set)



(a) S dataset

(b) A dataset

(c) F dataset

(d) G dataset

Fig. 7 Evaluation of different numbers of clusters for each set

specific data item (“what”). As discussed above, these “who” parameters are instances of the PPIoT *EntityType*. Therefore, we expect that clustering F permissions should provide an unanimous deny/share for all items, while clustering G permissions should

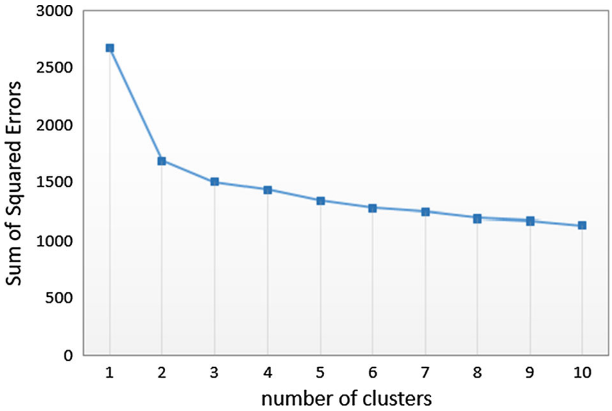


Fig. 8 Large error rates produced by directly clustering the entire dataset

provide more nuanced clusters of different entity types receiving the data specified in the F set.

6.2 Clustering methods

Our dataset shows considerable variability between participants' privacy preferences—a finding that is broadly reflected in the privacy literature (cf. Knijnenburg et al. 2013). Using clustering, one can capture the preferences of various users with a higher level of accuracy. Hence, the goal of this section is to find a concise set of profiles (clusters) that can represent the variability of the permission settings among our study participants.

To this end, we cluster participants' permissions with Weka[17] using the K-modes clustering algorithm (Chaturvedi et al. 2001) with default settings. The K-modes algorithm follows the same principles as the more common K-means algorithm, but it is more suitable for the nominal variables in our dataset.

In our first clustering attempt, we tried to find a set of profiles by clustering the entire dataset, including the S, A, F, and G subsets. A drawback of this method is that, assuming we cluster the participants into n clusters, this method will only provide n possible profiles to be used for recommendations to the users. A further drawback of clustering the full set of 45 permissions is that it gives large error rates for anything but a very large number of clusters (see Fig. 8; the sum of squared error for the viable 4-cluster solution is 1435).

If we instead generate a separate set of n "subprofiles" for each of the four datasets (S, A, F, and G), n^4 different combinations of profiles can be used for recommendation, providing finer-grained privacy-setting controls to the users compared to clustering the full set. In addition, error rates are lower when clustering each set separately, as shown in Fig. 7. For example, with only 2 clusters per set, the sum of squared error reduces to 1277 (a 24.3% reduction). A further benefit is that the profiles for each set can be investigated in more detail.

[17] https://www.cs.waikato.ac.nz/ml/weka/.

Another modeling decision must be highlighted as well: In our dataset, the fitness data permissions (F set) are specified repeatedly for each entity type (part of the G set). We tried to cluster these combinations, taking into account all 98 features (i.e., 14 fitness data per 7 entity types). This analysis resulted in two profiles: one that had “allow all” for health and SN public entities (and “deny all” for all other entities), and one that had “deny all” for all entities. This means that: a) very similar results can be obtained by considering the fitness data permissions separately from the entity type, and b) as expected, the “who” parameter (entity type) is more important than the “what” parameter (fitness data permissions).

In the following, we will discuss our method that generates subprofiles for each of the four datasets.

6.3 Clustering outcomes

We first investigate the optimal number of clusters by running the K-modes algorithm for 1–6 clusters with a 70/30 train/test ratio, using the sum of squared errors of the test set for evaluation. The results are shown in Fig. 7. Using the elbow method (Kodinariya and Makwana 2013), we conclude that 2 is the optimal number of clusters for each dataset[18].

The final cluster centroids of the 2-cluster solution for each dataset are shown in Fig. 9, together with the results of the 1-cluster solution. We describe the subprofiles of each set in the subsections below.

6.3.1 The *S* set

- **Minimal** (cluster A): This subprofile allows the minimum permissions needed to effectively run a fitness app. This includes identity, location, Bluetooth, motion & fitness, and mobile data permissions.
- **Unconcerned** (cluster B): This subprofile allows all permissions in this dataset.

6.3.2 The *A* set

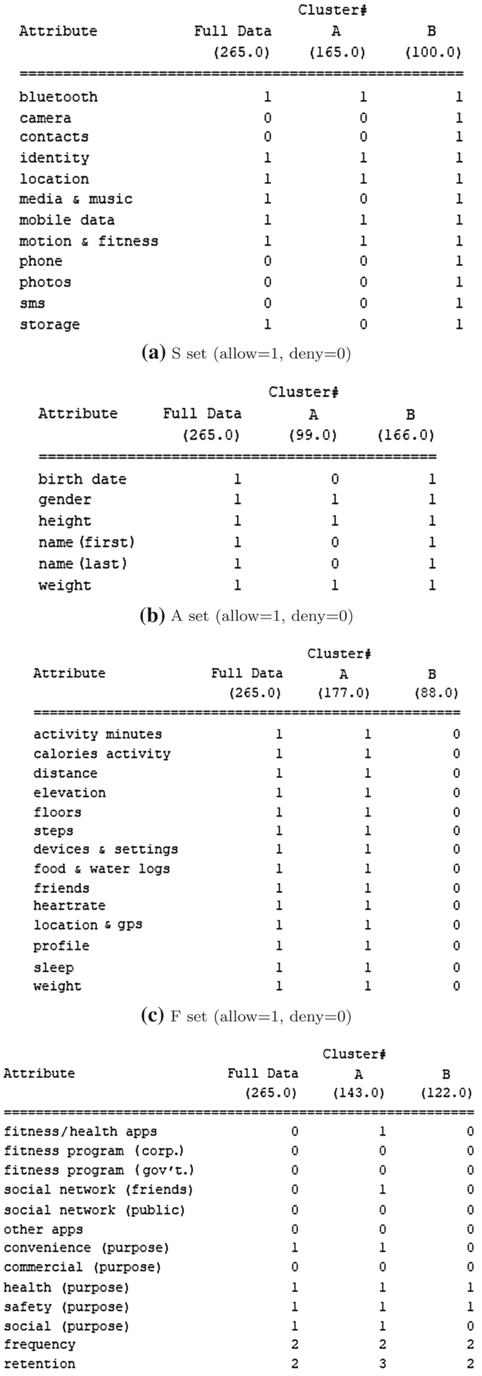
- **Anonymous** (cluster A): This subprofile shares only users’ gender, height, and weight information but not birth date or first and last name.
- **Unconcerned** (cluster B): This subprofile shares all data requested in this dataset.

6.3.3 The *F* set

- **Unconcerned** (cluster A): This subprofile shares all fitness data with TPs.
- **Strict** (cluster B): This subprofile does not share any fitness data with TPs.

[18] We obtain similar results using other clustering algorithms, such as hierarchical clustering.

Fig. 9 Privacy profiles from the two clustering methods: 1-cluster results (full data) and 2-cluster results (privacy subprofiles) for each dataset



```
                                      Cluster#
Attribute            Full Data          A            B
                      (265.0)        (165.0)      (100.0)
=====================================================
bluetooth                    1            1            1
camera                       0            0            1
contacts                     0            0            1
identity                     1            1            1
location                     1            1            1
media & music                1            0            1
mobile data                  1            1            1
motion & fitness             1            1            1
phone                        0            0            1
photos                       0            0            1
sms                          0            0            1
storage                      1            0            1
```

(a) S set (allow=1, deny=0)

```
                                Cluster#
Attribute      Full Data          A           B
                (265.0)        (99.0)     (166.0)
=============================================
birth date             1            0           1
gender                 1            1           1
height                 1            1           1
name(first)            1            0           1
name(last)             1            0           1
weight                 1            1           1
```

(b) A set (allow=1, deny=0)

```
                                        Cluster#
Attribute               Full Data          A           B
                         (265.0)        (177.0)     (88.0)
======================================================
activity minutes                1            1           0
calories activity               1            1           0
distance                        1            1           0
elevation                       1            1           0
floors                          1            1           0
steps                           1            1           0
devices & settings              1            1           0
food & water logs               1            1           0
friends                         1            1           0
heartrate                       1            1           0
location & gps                  1            1           0
profile                         1            1           0
sleep                           1            1           0
weight                          1            1           0
```

(c) F set (allow=1, deny=0)

```
                                               Cluster#
Attribute                     Full Data          A           B
                               (265.0)        (143.0)     (122.0)
===========================================================
fitness/health apps                   0            1           0
fitness program (corp.)               0            0           0
fitness program (gov't.)              0            0           0
social network (friends)              0            1           0
social network (public)               0            0           0
other apps                            0            0           0
convenience (purpose)                 1            1           0
commercial (purpose)                  0            0           0
health (purpose)                      1            1           1
safety (purpose)                      1            1           1
social (purpose)                      1            1           0
frequency                             2            2           2
retention                             2            3           2
```

(d) G set (allow=1, deny=0, except for frequency & retention)

Table 2 Association of cluster assignments for each pair of subsets, with odds ratio (OR) and p value

	A	F	S
F	OR: 1.34 $p = 0.266$	–	–
S	OR: 3.36 $p = 0.001$	OR: 2.56 $p = 0.001$	–
G	OR: 1.62 $p = 0.059$	**OR: 26.91** **p <.001**	OR: 2.54 $p < .001$

6.3.4 The *G* set

- **Socially active** (cluster A): This subprofile shares data with health/fitness apps and social network friends, but not with other recipients. Sharing is allowed for health, safety, and social purposes but not for commercial purposes.
- **Health-focused** (cluster B): This subprofile does not allow sharing with any TPs. Sharing is allowed only for health and safety purposes.

6.4 Cluster dependency

In Sect. 6.2, we argued for the creation of separate profiles for each of the four subsets of permissions (S, A, F and G sets). In this section, we validate this approach by testing the dependencies between the profiles of each pair of subsets using chi-square tests. If these tests are nonsignificant and/or show a small effect size, then the cross-subset dependencies between clusters are low, and it is indeed meaningful to cluster each subset independently.

The odds ratios in Table 2 indicate the dependencies between two particular subsets. For example, the odds of participants being clustered into cluster B of the F set were 1.34 times higher if they were clustered into cluster A of the A set than if they were clustered into cluster B of the A set. The odds ratios in Table 2 range from 1.34 to 26.91, and while all but one pair show a significant association, only one odds ratio represents a *substantial* association (i.e., a large[19] effect). The exception is the F and G set pair, with an odds ratio of 26.91 ($p < .001$). Coincidentally, the permissions of the F set and the G set are naturally connected in our design, since the permissions in the G set consider the disclosure of the permissions in the F set to “fourth-party” entities. Beyond this understandable association, we consider our clustering assignments per subset to be independent given above results.

7 Profile prediction

Now that we have identified two privacy “subprofiles” per dataset, the next step is to find predictors for the profiles and predict which subprofiles each participant belongs

[19] The generally accepted thresholds for odds ratios are 1.68 for a small effect size, 3.47 for a medium effect size, and 6.71 for a large effect size.

to. This section aims to answer the research question: **RQ3** Are there any privacy profile items or questionnaire items that can be used to predict which privacy profile best describes a user?

Recommender systems usually ask users to evaluate a few items before giving recommendations regarding all remaining items. Likewise, in our system, we might be able to identify certain permission items inside each privacy subprofile that—when answered by the user—could drive the prediction. Since the items are the permission preferences included in the subprofiles, collected through our FitPro prototype app, we call this the "direct prediction" approach. Additionally, we also explored whether the items from our questionnaire (see Sect. 5.2) could drive the prediction. Since these items are not part of the privacy subprofiles, we call this the "indirect prediction" approach. For each approach and for each subset of data (S, A, F, and G sets), we develop decision trees that will enable us to predict which subprofile best describes a user. The trees contain the subprofile items (direct prediction) or questionnaire items (indirect prediction) that can be asked to classify each user into their correct subprofile.

We developed our decision trees using the J48 tree learning algorithm. J48 is an efficient and widely used decision tree algorithm that can be used for classification (Patil and Sherekar 2013). Previous work shows the effectiveness of this approach to predict privacy settings within each cluster (Bahirat et al. 2018); here, we take the opposite approach and use it to predict cluster assignments instead. In our approach, the J48 algorithm extracts the permission items (for the direct prediction) or questionnaire items (for the indirect prediction) that classify a new user into the correct subprofile with the highest possible accuracy.

The tree results are reported in Table 3. For each determiner type, four trees are produced for A, S, F, and G sets, respectively. All trees produced binary leaves that output cluster A and B of each specific set. The condition of the cluster assignment is reported in the table. The number of assigned and incorrect predictions is also shown (i.e., #assigned/#errors) together with the prediction accuracy of each tree. The evaluation of all developed J48 trees was performed using k-fold cross-validation.

7.1 Direct prediction questions

In our direct prediction approach, the aim is to ask users to answer certain permission requests from each subset as a means to classify them into the correct subprofile (thereby providing a recommendation for the remaining items in that subset). For this approach, we thus classify users using the items in the subset as predictors.

Our results for this approach are reported in the "direct permission" column in Table 3. It shows, for each subset, the question (i.e., setting) that best classifies our study participants into the correct subprofile.

When running tree-based algorithms, a trade-off has to be made between the parsimony and the accuracy of the solution. Parsimony prevents over-fitting and promotes fairness (Bahirat et al. 2018) and can be accomplished by pruning the decision trees. In our study, while multi-item trees may provide better predictions, the increase in accuracy is not significant compared to the single-item trees presented in Table 3. These single-item solutions already obtained a high accuracy, and their parsimony

Table 3 Result of J48 trees for the permission drivers of the privacy subprofiles and their respective prediction accuracies

		Drivers							
		Direct permission		Privacy attitudes		Social behavior		Negotiability of privacy settings	
S Set	asked	Photograph		Privacy concern (question #4)		Sociability (question #2)		Phone (benefit increase)	
	(condition)	(if allowed)	(if denied)	(if > 4)	(if <= 4)	(if <= 1)	(if > 1)	(if allowed)	(if denied)
	prediction	Unconcerned	Minimal	Unconcerned	Minimal	Unconcerned	Minimal	Unconcerned	Minimal
	#assigned/#errors	72/4	193/32	74/30	191/56	25/8	240/83	107/39	158/32
	accuracy	86.42%		66.04%		65.66%		73.21%	
A Set	asked	First name		Trust (question #4)		Sociability (question #2)		Identity (risk decrease)	
	(condition)	(if allowed)	(if denied)	(if > 3)	(if <= 3)	(if <= 5)	(if > 5)	(if allowed)	(if denied)
	prediction	Unconcerned	Anonymous	Unconcerned	Anonymous	Unconcerned	Anonymous	Unconcerned	Anonymous
	#assigned/#errors	173/9	92/2	248/87	17/5	224/77	41/19	140/34	125/60
	accuracy	95.85%		65.28%		61.89%		62.26%	
F Set	asked	Activity minutes		Privacy concern (question #4)		Sociability (question #1)		Sleep (risk decrease)	
	(condition)	(if allowed)	(if denied)	(if > 1)	(if <= 1)	(if <= 6)	(if > 6)		(if denied)
	prediction	Unconcerned	Strict	Unconcerned	Strict	Unconcerned	Strict	(if allowed)	Unconcerned Strict
	#assigned/#errors	183/6	82/0	229/66	36/14	210/57	55/24	213/55	52/19
	accuracy	97.74%		69.81%		69.43%		72.08%	
G Set	asked	Social (GDPR purpose)		Trust (question #1)		Influence (question #2)		Profile (risk decrease)	
	(condition)	(if allowed)	(if denied)	(if > 5)	(if <= 5)	(if > 4)	(if <= 4)	(if allowed)	(if denied)
	prediction	Socially active	Health-focused	Socially active	Health-focused	Socially active	Health-focused	Socially active	Health-focused
	#assigned/#errors	142/23	123/24	128/53	137/47	125/49	140/46	154/50	111/39
	accuracy	82.26%		62.26%		61.89%		66.41%	

prevents over-fitting and minimizes the number of questions that will need to be asked to the users in order to provide them accurate recommendations. The resulting solution involves a 4-question input sequence—one question for each subset.

For the S set, the photograph permission is the best subprofile predictor. This is one of the least-shared permissions (see Fig. 5), and 94% of participants who give this permission are correctly classified into the "Unconcerned" subprofile, while 83% of participants who do not give this permission are correctly classified into the "Minimal" subprofile.

For the A set, first name is the best predictor. Again, 94% of participants who share their first name are correctly classified into the "Unconcerned" subprofile, while 98% of participants who do not share their first name are correctly classified into the "Anonymous" subprofile.

For the F set, the activity minutes permission is the best predictor. This is one of the most shared permissions. 97% of participants who give this permission are correctly classified into the "Unconcerned" subprofile, while 100% of participants who do not give this permission are correctly classified into the "Strict" subprofile.

Finally, for the G set, the best predictor is whether the participants allow data collection for social purposes. If so, participants are correctly classified into the "Socially active" subprofile with 84% accuracy, otherwise they are classified into the "Health-focused" subprofile with 80% accuracy.

7.2 Indirect prediction questions

A similar procedure was applied to the questionnaire data concerning the following categories of user traits: privacy attitude, social behavior, negotiability of privacy settings, exercise tendencies, and user demographics (cf. Table 4 in the Appendix).

User traits have been found to be associated with information disclosure and can hence be used to predict user privacy settings (Knijnenburg and Kobsa 2013; Knijnenburg et al. 2013; Li et al. 2017; Raber and Krüger 2018). For instance, Knijnenburg et al. (Knijnenburg and Kobsa 2013, Knijnenburg et al. 2013) show that users can be grouped according to their privacy attitudes, behaviors, and demographic characteristics. The best strategy for recommending users differs for each of these groups. Combining both privacy and personality measures (i.e., user traits) was shown to yield significant improvement in prediction accuracy in social networks for fine-grained location sharing (Raber and Krüger 2018). Finally, cultural traits significantly improve prediction accuracy, beyond demographics, and attitudinal and contextual factors (Li et al. 2017).

In the current study, we refer to the prediction of privacy preferences using user traits as *indirect prediction*. The indirect prediction approach has a lower accuracy than the direct approach presented in Sect. 7.1. This is expected, since unlike the direct prediction questions, the questionnaire items about user traits have no direct relationship with the permission settings in the privacy profiles. These results are still interesting, though, since they allow the user to avoid making any specific privacy settings [see Knijnenburg and Jin (2013) for a similar argument]. Moreover, the resulting predictors show interesting semantic relationships with the datasets they predict. We discuss these results in more detail in the following subsections.

7.2.1 Privacy attitude

We first attempted to use privacy attitude questions as predictors of users' subprofiles. The resulting trees for this indirect prediction approach are shown in the "privacy attitude" column of Table 3.

Among all the privacy attitude questions, "trust" and "privacy concern" are found to be predicting factors of user subprofiles. Interestingly, there is a single privacy concern question ("I believe other people are too concerned with online privacy issues") that predicts the user's S and F subprofiles. Those who agree that people are just too concerned about privacy issues belong to the "Unconcerned" subprofile, while those who have higher concerns tend to be in the "Minimal" subprofile. The same goes for the F set where those who strongly disagree—(1) on a 7pt scale, thinking that it is a major concern, belong to the "Strict" subprofile. Otherwise, they are classified as "Unconcerned."

For the trust question, "I believe the company is honest when it comes to using the information they provide," it can be used to predict users' subprofile for the A set. Participants are assigned to the "Anonymous" subprofile if they answer this question with "somewhat disagree" (3) or below. Those who indicate higher levels of trust are assigned to the "Unconcerned" subprofile. The A set concerns information provided directly to the fitness app, so it makes sense that trust is a significant predictor of users' willingness to provide such information.

For the G set, those users who agree (6) or extremely agree (7) with the question "I believe the company providing this fitness tracker is trustworthy in handling my information" are classified in the "Socially active" subprofile, while the remaining users are classified in the "Health-focused" subprofile. The question really fits the G set since GDPR permissions are mostly about handling the user information by the TPs. Particularly, it makes sense that users who do not trust the fitness app in handling their information would be assigned to the "Health-focused" profile, since this profile prevents the app from sharing their data to any other entity and only allows data collection for the purpose of health and/or safety.

The results highlight some semantically relevant relationships between users' attitudes and their assigned privacy profiles. The S and F sets share the same predictor question which makes the final solution a 3-question input sequence—this means that one fewer question must be asked, compared to the direct questions in Sect. 7.1.

7.2.2 Social behavior

We also searched for predictors among the questions about social influence and sociability. The resulting trees for this indirect prediction are shown in the "social behavior" column of Table 3.

A single sociability question can be used to predict subprofiles for both the S and A sets. For the S set, users who are completely open (1) to the idea of meeting new friends when they exercise are classified in the "Unconcerned" subprofile, otherwise they are classified in the "Minimal" subprofile.

For the A set, users who are likely not (6) or definitely not (7) open to meeting new friends are classified in the “Anonymous” subprofile, otherwise they are classified in the “Unconcerned” subprofile.

For the F set, users who have never (7) met any new friends while exercising are classified into the “Strict” subprofile, while others are classified into the “Unconcerned” subprofile. This, as well as the findings regarding the S and A sets, seems to suggest that users’ disclosure of personal information is likely to be related with their tendency to socialize while using fitness apps.

For the G set, users who are influenced to do exercise if their social media friends also exercise (i.e., “definitely yes” to “neutral” (1–4)) are classified into the “Socially active” subprofile, otherwise they are classified into the “Health-focused” subprofile.

Again, we found interesting semantic relationships between social influence and sociability while exercising and users’ privacy-related behaviors: Users who are more prone to reap social benefits from exercising are more likely to give the app more widespread permissions. Similarly to privacy attitudes, these predictors only involve a 3-question input sequence.

7.2.3 Negotiability of privacy settings

We also attempted to use the negotiability of users’ privacy settings as input for subprofile prediction. The “negotiability of privacy settings” column of Table 3 shows the tree learning solutions for this approach.

For the S set, users who are willing to give the phone permission (access phone calls and call settings) if the benefits increase are classified into the “Unconcerned” subprofile, while users who refuse to share the phone permission even if the benefits increase are classified into the “Minimal” subprofile. In other words, the privacy preferences of the latter group are not negotiable; they will still share only the minimum permissions needed to run the tracker, even if the benefits increase.

For the A set, users who are willing to give the identity permission (account and/or profile information) if the risks decrease are classified into the “Unconcerned” subprofile, otherwise they are classified into the “Anonymous” subprofile. Interestingly, the identity permission is part of the S set rather than the A set, but it semantically coincides with the items in the A set, which include the user’s name and birth date (i.e., identifying information). As such, it makes sense that users who are unwilling to share their phone’s identifier even when the risks decrease are also unwilling to share their personal identity information.

For the F set, users who share their sleep fitness data with other TPs if the risks decrease are classified into the “Unconcerned” subprofile, otherwise they are classified into the “Strict” subprofile. Users in the latter subprofile will not share their fitness data with any other TPs, even if the risk decreases.

For the G set, users who share their fitness app profile with other TPs if the risks decrease are classified into the “Socially active” subprofile, otherwise they are classified into the “Health-focused” subprofile. Even though profile is a permission from the F set, it semantically coincides with the subprofiles of the G set: Users in the “Socially active” subprofile tend to have permissions that allow them to connect to others while exercising, and sharing one’s fitness app profile is indeed a potential way to connect

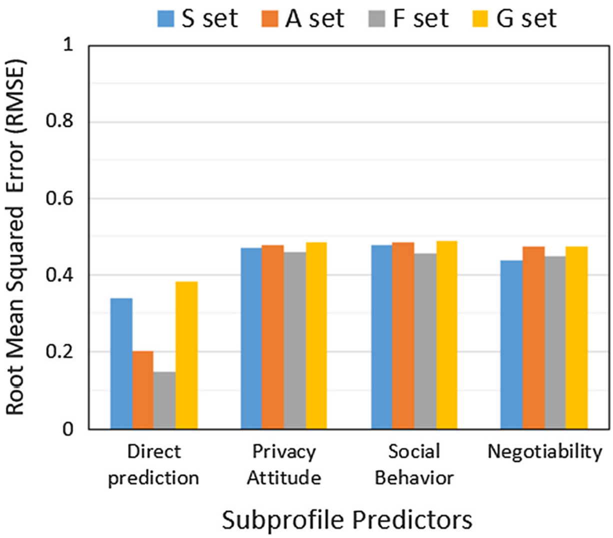


Fig. 10 Tree evaluation. Root-mean-square error for each J48 tree algorithm

to other users. As such, it makes sense that users in this subprofile are more willing to share their fitness app profile if the risks of doing so decrease.

The classification accuracy of the negotiability questions is the highest among all "indirect prediction" approaches. The most predictive questions also have understandable semantic relationships with the datasets they predict.

7.2.4 Exercise tendencies and user demographics

We applied tree learning algorithms to the group of exercise tendency questions and user demographics as well, but we found no significant predictors among these questions. While other studies have found user demographics to be significant predictors of privacy behaviors (Knijnenburg and Kobsa 2013), in this particular study we were not able to find any significant predictors among the group of user demographics.

7.3 Tree evaluation

Figure 10 shows the root-mean-square error of all the trees produced by the J48 classifier. The evaluation has been executed with k-fold cross-validation with $k = 10$.

As expected, the "direct prediction" approach results in lower error rates than the various "indirect prediction" approaches, since in the former approach the items are a direct part of the privacy settings that constitute the subprofiles. Among the "indirect prediction" approaches, the *negotiability of privacy settings* has slightly lower error rates. This is not surprising, since it is at least partially related to the privacy settings (yet evaluates whether those settings will change under certain conditions).

8 Recommendation strategies and validation

In this section, we describe different types of guided privacy-setting approaches that are based on the previous clustering and tree learning results. When implemented in the PDM, the guided interface simplifies the privacy-setting experience by providing privacy recommendations. This answers **RQ4**: How can we effectively exploit the results to provide recommendation? We also present a validation of the recommendation results using a holdout sample of permission settings from 30 additional users. The PDM design prototype implementing the recommendation strategies is available online[20].

8.1 Privacy-setting recommendations

8.1.1 Manual setting

The baseline privacy settings interface is one where users have to manually set their settings, which are initially turned off. If users do this correctly, these manual settings should match their privacy preferences 100%. However, the process of manually setting one's privacy settings can be very burdensome for the user; our system has a total of 45 permissions that are required to be managed. Under such burden, users are likely going to make mistakes (cf. Madejski et al. 2012), so the 100% accuracy may not be achieved through manual settings.

The next strategies exploit the results of the analysis in the previous section to provide *interactive recommendations* that simplify the task of privacy permission setting, with different levels and type of user intervention.

8.1.2 Single smart default setting

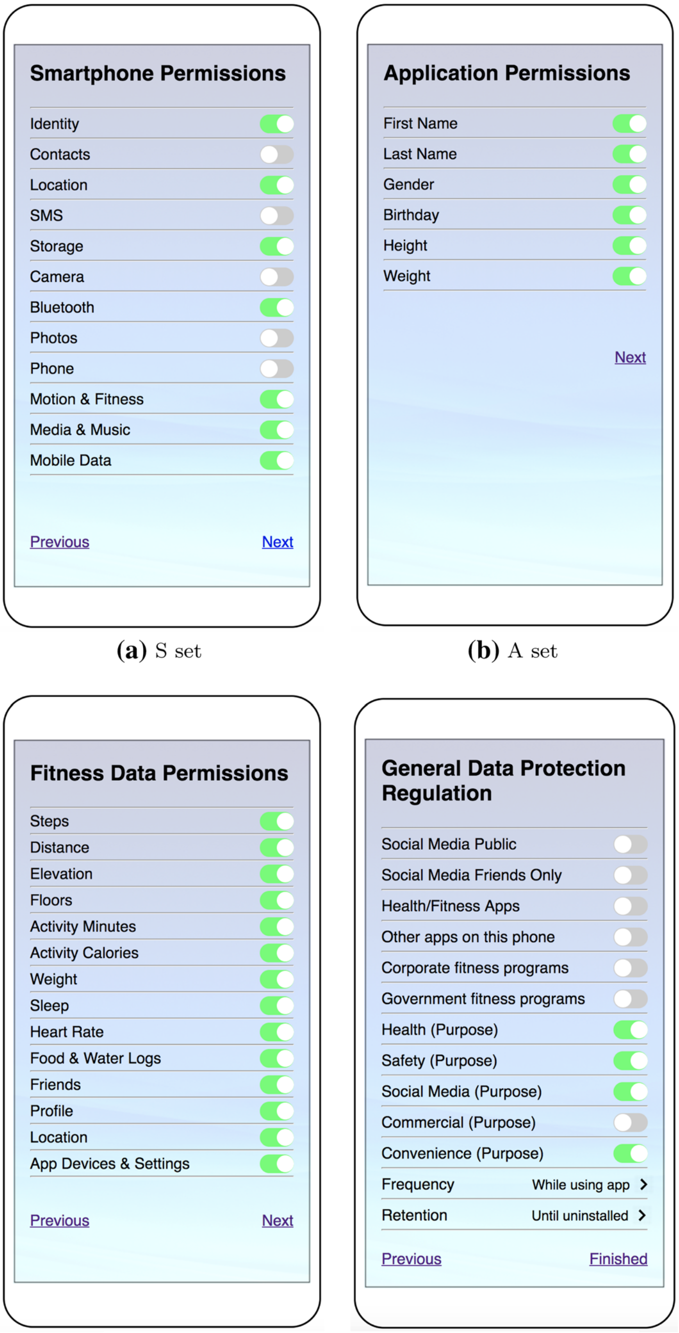
One way to reduce the burden of privacy management is with single "smart" default setting. Rather than having the user set each permission manually, this solution already selects a default setting for each permission. Users can then review these settings and change only the ones that do not match their preferences.

The optimal "smart" default is a set of settings that is aligned with the preferences of the majority of users. Hence, we can calculate these settings by using the cluster centroid of the 1-cluster solution (i.e., the "full data" column in Fig. 9). Figure 11 shows the resulting default values for each dataset. If the user is unhappy with these settings, she/he can still make specific changes. Otherwise, he/she can keep them without making any changes.

8.1.3 Pick subprofiles

The single smart default setting works best when most users have preferences similar to the average. However, our dataset shows considerable variability in participants' privacy preferences—a finding that is broadly reflected in the privacy literature (cf.

[20] The UI design can be found in http://pdm-aids.dibris.unige.it/interface/.



Smartphone Permissions

- Identity ☑
- Contacts ☐
- Location ☑
- SMS ☐
- Storage ☑
- Camera ☐
- Bluetooth ☑
- Photos ☐
- Phone ☐
- Motion & Fitness ☑
- Media & Music ☑
- Mobile Data ☑

(a) S set

Application Permissions

- First Name ☑
- Last Name ☑
- Gender ☑
- Birthday ☑
- Height ☑
- Weight ☑

(b) A set

Fitness Data Permissions

- Steps ☑
- Distance ☑
- Elevation ☑
- Floors ☑
- Activity Minutes ☑
- Activity Calories ☑
- Weight ☑
- Sleep ☑
- Heart Rate ☑
- Food & Water Logs ☑
- Friends ☑
- Profile ☑
- Location ☑
- App Devices & Settings ☑

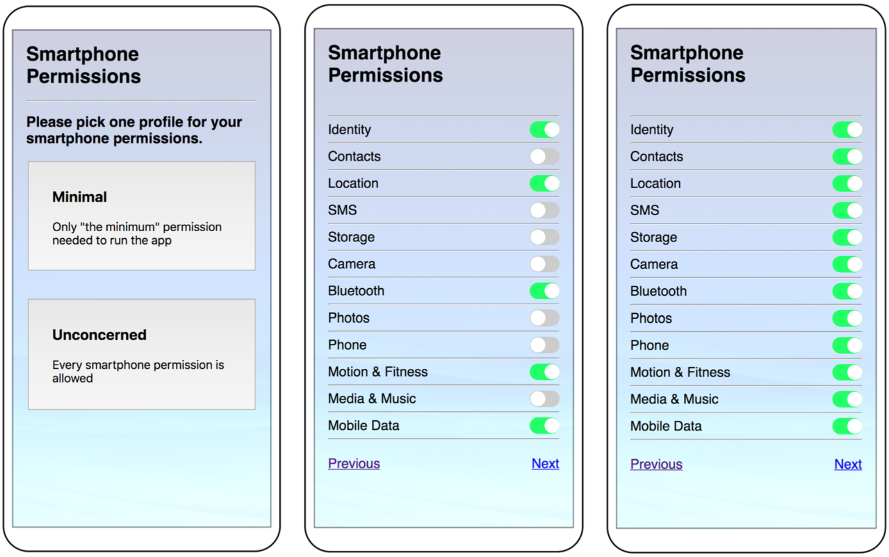
(c) F set

General Data Protection Regulation

- Social Media Public ☐
- Social Media Friends Only ☐
- Health/Fitness Apps ☐
- Other apps on this phone ☐
- Corporate fitness programs ☐
- Government fitness programs ☐
- Health (Purpose) ☑
- Safety (Purpose) ☑
- Social Media (Purpose) ☑
- Commercial (Purpose) ☐
- Convenience (Purpose) ☑
- Frequency: While using app >
- Retention: Until uninstalled >

(d) G set

Fig. 11 Smart single settings



Smartphone Permissions

Please pick one profile for your smartphone permissions.

Minimal
Only "the minimum" permission needed to run the app

Unconcerned
Every smartphone permission is allowed

Smartphone Permissions

Identity ●
Contacts ○
Location ●
SMS ○
Storage ○
Camera ○
Bluetooth ●
Photos ○
Phone ○
Motion & Fitness ●
Media & Music ○
Mobile Data ●

Previous Next

Smartphone Permissions

Identity ●
Contacts ●
Location ●
SMS ●
Storage ●
Camera ●
Bluetooth ●
Photos ●
Phone ●
Motion & Fitness ●
Media & Music ●
Mobile Data ●

Previous Next

(a) S set subprofiles **(b)** The "Minimal" subprofile **(c)** The "Unconcerned" subprofile

Fig. 12 Interaction for picking a subprofile for the S set

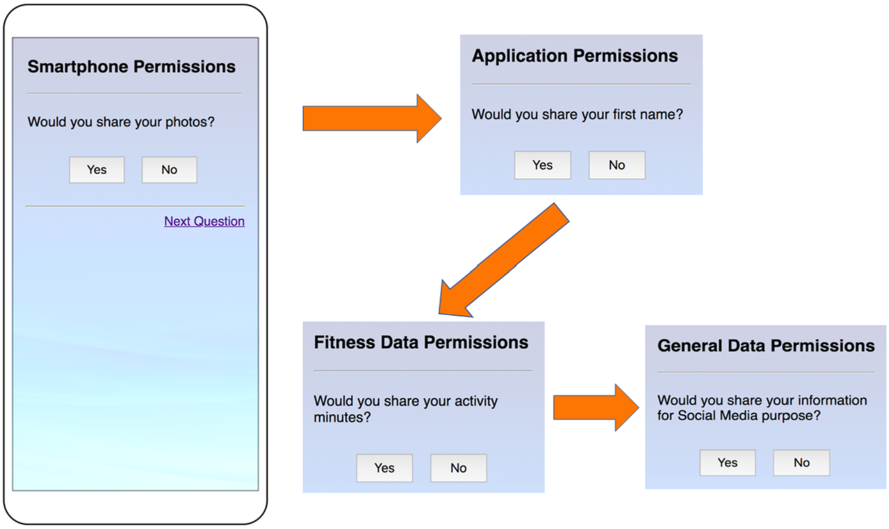
Knijnenburg et al. 2013). This brings us to our clustering solutions, which create *separate* default settings (in the form of subprofiles) for distinct groups of users.

Our first approach in this regard is to have users manually select which privacy subprofiles they prefer. Figure 12a shows the subprofile selection interface for the S set. Users can choose either the "Minimal" or "Unconcerned" subprofile, which are shown in Figure 12b, c, respectively. Similar interfaces are provided for the A, F, and G sets (not shown here).

The subprofiles provided by this approach have a higher overall accuracy than the single "smart" default described in Sect. 8.1.1, meaning that the user will have to spend less effort changing the settings. However, the user *will* have to select a subprofile for each dataset. This highlights the importance of having a small number of subprofiles and making these subprofiles easy to understand. That said, even with only two subprofiles per dataset, this can be a challenging task. In the next two subsections, we address this problem by automatically selecting subprofiles based on users' answers to specific subprofile items ("direct prediction") or questionnaire items ("indirect prediction").

8.1.4 Direct prediction

For the direct prediction approach, we devise an interactive 4-question input sequence as shown in Fig. 13. Each question asks users' decision on a specific permission, which guides the subprofile classification processes as outlined in Sect. 7.1. In effect, each question informs the system about the user's subprofile of one of the four datasets, which means that users no longer have to manually pick the correct subprofiles. Specifically, users will be asked if they agree to share their photographs (for the S set



Smartphone Permissions

Would you share your photos?

Application Permissions

Would you share your first name?

Fitness Data Permissions

Would you share your activity minutes?

General Data Permissions

Would you share your information for Social Media purpose?

Fig. 13 Direct prediction questions

recommendation), first name (for the A set recommendation), activity (for the F set recommendation), and whether they allow their data to be used for Social purposes (for the G set recommendation). This 4-question interaction will aid the users in setting all of the 45 permissions in the system. Depending on the answer to these questions, the user will subsequently see the settings screens with the defaults set to the predicted profile. Users can still change specific settings if their preferences deviate from the selected profile.

8.1.5 Indirect prediction

Similarly, an interactive 4-question input sequence is created to collect users' privacy-setting preference for the indirect prediction approach. Compared to the direction prediction approach, these 4 questions are selected from the questionnaire items instead of the permission settings. The questions are selected in a manner that yields the highest accuracy for each permission set: a negotiability question for phone permissions for the S set (*Would you share your phone permission if the benefits increase?*), a question about sociability for the A set (*Are you open to the idea of meeting new friends while you exercise?*), a negotiability question for the permission to share sleep data for the F set (*Would you share your sleep permission if the risks decrease?*), and a trust question for the G set, (*Do you believe that the company providing this fitness tracker is trustworthy in handling your information?*). Negotiability and attitude have almost the same accuracy for G set, so we chose attitude for the sake of question diversity.

The answers to these questions are used to automatically recommend corresponding setting to users. The benefit of the indirect prediction approach is that the user does not have to answer any permission questions, not even the four needed in the direct

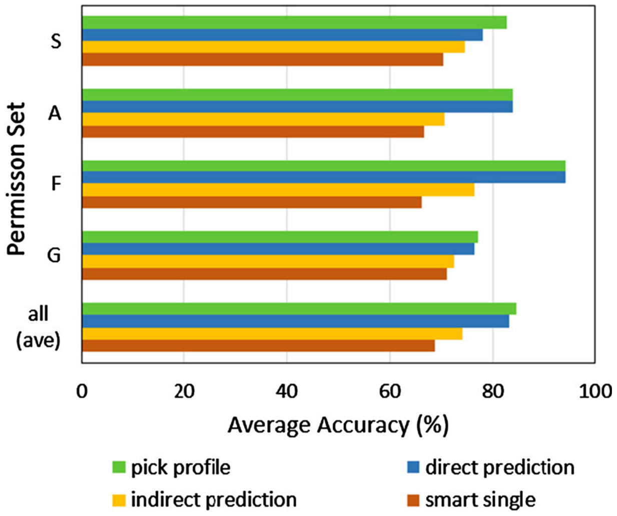


Fig. 14 Average accuracies of the recommender strategies on the 30 users in our holdout dataset

approach to give them a subprofile recommendation. Instead, they just need to answer four questionnaire items.

8.2 Validation

We conducted a validation of these different approaches by running the recommendation strategies on the 30 users in our holdout dataset. The resulting recommended privacy subprofiles are then compared with their actual privacy preferences. Figure 14 shows the average accuracies of each of the presented approaches.

The *Pick Profile* approach reaches an 84.74% accuracy. This approach has the highest accuracy, because only the error from the difference between the privacy profile and the users' settings is counted, omitting the errors introduced by the user classification. This assumes that users can classify themselves with perfect accuracy—this is likely an incorrect assumption.

Among recommendation approaches, the *direct prediction* approach is the most accurate, averaging 83.41%. It almost yields no additional classification error compared to the *Pick subprofile* approach. The *indirect prediction* approach has a significantly lower accuracy of 73.9%.

Finally, the *single smart default* approach uses only a single "profile," circumventing the need for classification. The default profile settings are shown in the "full data" column of Fig. 9. The accuracy of this setting is lower than the accuracy of the subprofile solutions, but it does not lose accuracy on classification. Hence, its accuracy is a respectable 68.7%, which is not much lower than the *indirect prediction* approach.

The details about accuracies are provided in Table 5 in the Appendix.

9 Discussion, limitations, and future work

9.1 Advanced privacy recommendations

Unsupervised machine learning methods like clustering do not have a ground truth, and the optimal number of clusters is therefore a subjective decision that can depend on many factors. Theoretically, an error-free solution can be attained when the number of clusters is equal to the number of distinct settings—a solution that does not generalize well because it is over-fitting the data. In this study, we therefore followed the elbow method: We picked the number of clusters where the within-cluster sum of squared error transitions from a sharp decrease to a negligible decrease. This approach is justified in our validation (see Sect. 8.2), as it yields an overall accuracy of 84.74%. A larger number of clusters will likely somewhat increase this accuracy, but this would come at the cost of complexity.

From a machine learning perspective, our 2-cluster solution per data subset is rather simple. It is, however, justified, for two reasons. First, our analysis of the optimal number of clusters using the elbow method shows that two is clearly the optimal solution for each subset—a solution with more clusters would not be substantially more accurate. Second, our goal in this endeavor is not to get the most interesting or even the most accurate clustering result, but to help end-users by simplifying their privacy decisions. Two profiles per subset are the least complicated solution, which is arguably a benefit for our system, especially in the scenario where we ask users to manually choose among the resulting profiles. In the end, then, we believe that this is a trade-off between model accuracy and parsimony, which should be explored further in future work.

We note that our results are by no means trivial. Indeed, by evaluating the clustering for each set of permissions separately, we end up with a total of 16 (2^4) subprofile combinations. The dependency of the clusters of the different subsets is low (see Sect. 6.4), which justifies our per-subset approach. Moreover, while the two profiles of our subsets tend toward one of the extremes, only the F set has “all on” and “all off” as its two profiles—the others are more sophisticated. These somewhat more trivial profiles nonetheless yield an accuracy of 94.29%. All accuracy measures are reported in Table 5 in the Appendix.

In this light, it is important to note that users are still given the option to make manual changes to their privacy settings after a profile is selected. So even if some of these profiles tend toward one of the extremes, we argue that they are mostly meant to be a helpful starting point closest to the users real preferences. We see their simplicity as an advantage, as they are easy to comprehend, unlikely to overfit our particular data, and likely to generalize to other scenarios.

The profiles, then, are a convenient *shortcut* to help users with their privacy decisions. While researchers have long argued that users make carefully considered decisions regarding their data privacy (cf. by applying a “privacy calculus” Dinev and Hart 2006), empirical work has shown that people rarely take the time and effort to carefully weigh the risks and benefits of their decisions as the privacy calculus suggests. Instead, many privacy decisions are heuristic and thereby subject to a large number of decision fallacies (Knijnenburg et al. 2017). For example, research on the "default effect" (Johnson et al. 2002) shows that people tend to follow default settings in making decisions. As such, when all settings are off by default, they tend to end up sharing less than when all settings are on by default. Practically speaking, our profiles represent the default setting that is closest to users' actual preferences, thereby reducing the default effect.

More generally, our privacy recommendation procedure uses the privacy calculus as a "prescriptive" model: Instead of burdening users with the task of balancing privacy and benefit, our recommendation strategy provides users with a guidance to the privacy calculus, based on an analysis of the decisions of a large number of other users (which is used to generate the profiles and the profile-assignment strategy). As such, the value of this proposed procedure extends beyond the specific profiles discussed in this paper.

Our quest for simplicity does not mean that our recommendation strategies cannot be improved. For example, one of the limitations of our recommendation strategies is that they are static: A potential improvement would be to update the recommendations automatically based on new input. Such a dynamic recommender would have some drawbacks, though: If the recommender is to update predictions for the *current user* based on their feedback, it has only very limited opportunities to do so, since the interaction consists of only four screens (unlike a typical recommender system, where users have continual interactions with the system). Likewise, if the recommender is to learn from each user and recalculate the recommendations for *subsequent users*, it means that the system needs some sort of centralized learning component where all users' privacy preferences are stored. This in itself requires that users give permission for their privacy preferences to be stored and processed.

To summarize, our aim in this paper is to study which tracking data are viable for determining the right recommendation in a simplified manner. For future refinements, we plan to use dynamic cognitive environment techniques (e.g., dynamic Bayesian filtering, Kalman filtering, PhD filtering, etc.) that provide update steps to extend our static approach. Moreover, we plan to combine direct recommendation and indirect recommendation, which are two different strategies that result from our current studies.

9.2 Dataset limitations

Our dataset is collected via crowdsourcing using a simulation environment with a mockup system. Crowdsourcing has become an established mechanism in academia and industry to gather rich user study feedback (Zhao and Zhu 2014), user preferences and ideas regarding new product designs (Schemmann et al. 2016), quantitative user experiment data (Joosse et al. 2015), and input data for machine learning studies (Abhigna et al. 2018). Compared to in-lab studies, it tends to provide access to more culturally diverse samples, allowing for cross-cultural comparisons (cf. Joosse et al. 2015; Li et al. 2017), or narrowing down to a specific set of users.

The main downsides of crowdsourcing are the variable quality of worker input (which can be mitigated through proper compensation, time monitors and attention check questions—we applied all of these mechanisms, resulting in a reduction in our sample from 310 to 295 participants with valid data) and the lack of realism that comes with a simulated study environment. While we acknowledge that our simulated approach (with a system mockup) lacks some ecological validity, we also note that several other studies have used simulations or even scenarios (which forego a mockup altogether) since it is a more flexible and convenient way to gather feedback from potential users (Bahirat et al. 2018; He et al. 2019). Note that even in a field trial, our system would likely have to be a mockup, since none of the available fitness trackers implement all of the privacy settings we consider (e.g., the GDPR set). Choosing a simulated environment is therefore a viable forward-looking solution to measure users' privacy preferences regarding these new settings.

It is possible that our participants would have been biased by the fact that our study was specifically simulating the privacy-setting experience of an application. To reduce the emphasis on privacy, we avoided mentioning the term "privacy" throughout the study. Moreover, to limit the differences between our simulation and an actual app installation, we made the interaction design and the user interface of the app very realistic, and we asked users to behave like how they usually install an app. In this light, we note that Lee and Kobsa (2017) recently studied privacy-setting using both a Web-based survey and a Google Glass field trial. The results were only slightly different between these two study procedures, and the differences were more likely due to a difference in the recruited sample than in the realism of the study setup. This finding is echoed in the human–robot interaction domain, where simulated video-based trials are often used as a proxy for live interactions with an actual robot. Several works have shown that this method produces accurate results (Walters et al. 2011; Woods et al. 2006). As such, we are rather confident that our simulated environment captures users' real behaviors. That said, we suggest that a field study would nonetheless be an interesting endeavor for future work.

Finally, while we propose several recommendation strategies in this paper, we have not tested their operational efficacy from the user's perspective. Specifically, we have conjectured that profile-based approaches reduce the hassle of making privacy settings but that the manual selection of a privacy profile might be difficult for a user. These conjectures should be evaluated in a user study, which is another suggestion for future work. This user study could also evaluate the amount of user control that should be provided by the PDM.

9.3 Extending the PPIoT vocabulary

The vocabulary for the fitness data privacy model is based on the PPIoT ontology that we designed within our framework for personal data management in the IoT (Sanchez et al. 2019; Torre et al. 2016c, 2018). PPIoT has been designed following the well-established methodology in Noy et al. (2001) and evaluated accordingly using the competency questions method. Moreover, its logical consistency has been evaluated using the Jena Semantic Web reasoner, and its effectiveness has been tested in a task-based validation (cf. Brank et al. 2005; Hlomani and Stacey 2014), where it was used for users' privacy preference representation and matched against TP privacy statements. The demonstrator can be found online[21]. We are currently working on extending the

[21] https://github.com/OdnanOriginal/PDM

use case scenarios in order to comprehensively evaluate the PDM's feasibility to model the user and TP privacy preferences in the domain of IoT personal devices.

With respect to GDPR permissions, we acknowledge that our current model includes the *entity type* for data sharing and the *purpose*, but not the *means* of personal data processing. This is because it is difficult for end-users to understand the implications of various means of data processing and related issues (e.g., the meaning of encryption, differential privacy, etc.). While this would make it a good candidate for decision support, it also means that it is very difficult to get an accurate understanding of users' preferences regarding various means of data processing. We plan to carefully study this issue in future work, at which point we can address it in an update to our PPIoT vocabulary.

9.4 Extending the scope of our work

While our study focuses on fitness trackers, our work can easily be applied to fitness devices beyond wearables (e.g., mobile phones, smart cardio devices in the gym), as the settings and even the setting interfaces would likely be very similar. We also note that the presented approach in this paper has successfully been employed for public IoT (Bahirat et al. 2018) and household IoT settings (He et al. 2019). One way the current paper goes beyond this previous work is by explicitly considering GDPR permissions; this aspect can easily be integrated into this previous work as well, since it is applicable to any data-accessing entity, not just fitness-based applications.

Our work extends the scope of data collection to what one might call "fourth parties" that may use the data collected by a fitness tracking app. This aspect has not received a lot of attention in previous work but is relevant in ours, given the GDPR mandate regarding the free movement of personal data (The European Parliament and the Council of the European Union 2016). Our approach accounts for this aspect; however, we admit that it can be further investigated and improved. For example, given the GDPR notion of free movement of personal data, one could consider a fitness device as a platform that is not necessarily linked to a specific provider (i.e., a dedicated "third party") but that can instead be accessed by any application, pending the user's permission. For our future work, we will focus on this aspect more closely, especially given that major recent privacy breaches have occurred in complex, interdependent bundles of services (e.g., Facebook and Cambridge Analytica). Our work can be extended by further studying the complex personal data sharing ecosystems of modern third/fourth parties (Conger et al. 2013; Kurtz 2018).

10 Conclusion and contribution

In this paper, we presented a data-driven approach to the development of recommendation strategies for supporting users to set permissions regarding their personal data collected and shared by tracking devices in the fitness domain.

The motivating issue is the complex scenario of data sharing among devices and third party (TP) applications in the IoT, which makes setting one's privacy preferences an increasingly complex task. The goal is to balance the users' control over their data and the simplicity of setting, especially in light of GDPR requirements.

First, we defined a *data model* of *privacy preferences* for the *fitness domain* that can be represented using the vocabulary of our PPIoT ontology. The data model is based on the superset of permissions required by the most popular fitness trackers and includes the permissions specified in the GDPR. The use of a vocabulary aims to provide an unambiguous and formal representation of the user's privacy preferences, regardless of the diverse representations used by the TPs themselves. The PPIoT vocabulary is part of our personal data manager (PDM) framework which is the intended testbed for the privacy preference recommendation strategies that we propose in this paper.

Despite the vast variation in user privacy preferences, we managed to find a concise set of relevant privacy profiles that are able to represent these preferences. With two subprofiles for each of four subsets of permissions (sets S, A, F and G in the paper), a total of 16 possible privacy profiles can be recommended to the user. Additionally, we managed to determine specific subprofile items ("direct prediction") and questionnaire items ("indirect prediction") that serve as predictors for these profiles.

Our results also show interesting semantic relationships between predictors and privacy settings. In particular, users' tendency to make friends while using the fitness tracker is a significant predictor of the fact that they accept smartphone data permission requests (the S set), answer in-app requests (the A set), and share their fitness tracking data (the F set).

This study also found that in sharing fitness tracking data, users care more about "who" will receive that data rather than "what" data are shared specifically. This confirms previous studies (Bahirat et al. 2018; Lee and Kobsa 2016) showing no significant interaction between these two parameters. Our results also show that knowledge about users' actions when risks decrease is more useful to give good recommendations than knowledge about users' actions when benefits increase.

Finally, we proposed different recommendation strategies and related user interfaces for supporting users to set their privacy permissions. They include a fully manual approach, as well as interactive prediction-based recommendations that are based on our clustering and classification results. Users can interact with the user interface by answering the "trigger questions" that are selected by our classifiers as predictors of users' subprofiles. These recommendation approaches are aligned with the PPIoT vocabulary: The data model and the recommendation strategies will be used by the PDM to represent the user privacy preferences and recommend privacy settings.

Even though several works exist on privacy preference modeling, this paper makes a contribution in modeling privacy preferences for data sharing and processing of tracked data in the IoT and fitness domain, with specific attention to GDPR compliance. Moreover, the identification of well-defined clusters of preferences and predictors of such clusters is a relevant contribution for the design of recommendation strategies and interactive user interfaces that aim to balance users' control over their privacy permissions and the simplicity of setting these permissions.

In this light, our main contribution is a generic method to develop user profiles and a series of recommendation strategies for privacy management that can be applied to any user-tailored privacy decision support system that models and manages users' privacy permissions, like our PDM. Beyond existing work, we not only develop privacy profiles, but also identify potential predictors of these profiles. Such predictors include privacy setting preferences (direct prediction) but also, and more interestingly, some user traits (indirect prediction): users' privacy attitudes, the negotiability of their preferences, and social behavior.

With the limitations discussed in the previous section, our results could be immediately integrated in personalized services in the fitness domain. In fact, our data model for the fitness domain has a wide coverage of tracking data that likely include those used by existing personalized fitness services.

As argued, though, this approach can also be applied to other IoT scenarios (e.g., household IoT, public IoT), or even other complex privacy situations (e.g., social networking, online shopping) as well. We encourage researchers to adopt and further extend this "User-Tailored Privacy" approach (cf. Knijnenburg 2017) in their own work.

Appendices

Table 4 Study Questionnaire

	Privacy-related attitude questions (7pt scale)
Trust	I believe the company providing this fitness tracker is trustworthy in handling my information
	I believe this company tells the truth and fulfills promises related to the information I provide
	I believe this company is predictable and consistent regarding the usage of my information
	I believe this company is honest when it comes to using the information I provide
General privacy concerns	All things considered, the Internet causes serious privacy problems
	Compared to others, I am more sensitive about the way online companies handle my personal information
	To me, it is the most important thing to keep my privacy intact from online companies
	I believe other people are too concerned with online privacy issues
	Compared with other subjects on my mind, personal privacy is very important
	I am concerned about threats to my personal privacy today
Perceived surveillance	I believe that the location of my mobile device is monitored at least part of the time
	I am concerned that mobile apps are collecting too much information about me
	I am concerned that mobile apps may monitor my activities on my mobile device

Table 4 continued

	Privacy-related attitude questions (7pt scale)
Perceived intrusion	I feel that as a result of my using mobile apps, others know about me more than I am comfortable with
	I believe that as a result of my using mobile apps, information about me that I consider private is now more readily available to others than I would want
	I feel that as a result of my using mobile apps, information about me is out there that, if used, will invade my privacy
Perceived secondary use of personal information	I am concerned that mobile apps may use my personal information for other purposes without notifying me or getting my authorization
	When I give personal information to use mobile apps, I am concerned that apps may use my information for other purposes
	I am concerned that mobile apps may share my personal information with other entities without getting my authorization
	Negotiability of privacy settings questions (Y|N for each permission setting)
	Would you share the following data if the risks significantly increased?
	Would you share the following data if the benefits significantly decreased?
	Would you share the following data if the risks significantly decreased?
	Would you share the following data if the benefits significantly increased?
	Social behavior questions (7pt scale)
Social influence	If your friends exercise, does this influence you to exercise?
	If your social media friends exercise, does this influence you to exercise?
Sociability	How often do you meet new friends while you exercise?
	Are you open to the idea of meeting new friends while you exercise?
	Exercise tendencies questions (7pt scale; multiple choice for What questions)
Exercise attitude	How physically healthy are you?
	How important is exercise to you?
	What do you most often do for exercise?
	How often do you exercise?
	At what intensity do you work out?
	Do you feel you get too much, the right amount, or too little exercise?
	What is the main reason you exercise?
Healthy living expertise	I understand difference between different types of healthy-living measures
	I know healthy-living measures that most others haven’t even heard of
	I know which healthy-living measures are useful to implement
	I am able to choose the right healthy-living measures

Table 5 Table of accuracies

	Pick profile (%)	Single smart default (%)	Direct prediction (%)	Privacy attitude (%)	Social behavior (%)	Negotiability (%)
S Set						
Identity	66.67	66.67	66.67	66.67	66.67	66.67
Contacts	83.33	70.00	70.00	56.67	73.33	80.00
Location	83.33	83.33	83.33	83.33	83.33	83.33
SMS	90.00	50.00	70.00	50.00	53.33	73.33
Storage	83.33	56.67	70.00	43.33	46.67	60.00
Camera	80.00	60.00	86.67	60.00	70.00	63.33
Bluetooth	83.33	83.33	83.33	83.33	83.33	83.33
Photographs	80.00	66.67	100.00	60.00	76.66	70.00
Phone	96.67	56.67	76.67	50.00	60.00	80.00
Motion	96.67	96.67	96.67	96.67	96.67	96.67
Media	70.00	76.67	56.67	43.33	33.33	60.00
Mobile Data	76.67	76.67	76.67	76.67	76.67	76.67
Average	82.50	70.28	78.06	64.17	68.33	74.44
A set						
First Name	100.00	63.33	100.00	63.33	73.33	56.67
Last Name	96.67	60.00	96.67	60.00	70.00	60.00
Gender	76.67	76.67	76.67	76.67	76.67	76.67
Birthday	90.00	60.00	90.00	60.00	63.33	53.33
Height	70.00	70.00	70.00	70.00	70.00	70.00
Weight	70.00	70.00	70.00	70.00	70.00	70.00
Average	83.89	66.67	83.89	66.67	70.55	64.44
F set						
Steps	96.67	73.33	96.67	76.67	70.00	76.67
Distance	96.67	73.33	96.67	76.67	70.00	76.67
Elevation	100.00	70.00	100.00	73.33	73.33	80.00
Floors	96.67	73.33	96.67	76.67	70.00	76.67
Activity minutes	100.00	70.00	100.00	73.33	73.33	80.00
Calories activity	96.67	73.33	96.67	76.67	70.00	76.67
Weight	90.00	60.00	90.00	63.33	70.00	76.67
Sleep	93.33	63.33	93.33	66.67	66.67	80.00
Heart rate	100.00	70.00	100.00	73.33	73.33	80.00
Food logs	90.00	60.00	90.00	63.33	70.00	76.67
Friends	83.33	53.33	83.33	56.67	63.33	70.00
Profile	96.67	66.67	96.67	70.00	76.67	76.67
Location	86.67	56.67	86.67	60.00	66.67	66.67
Device & settings	93.33	63.33	93.33	66.67	73.33	73.33
Average	94.29	66.19	94.29	69.52	70.48	76.19

Table 5 continued

	Pick profile (%)	Single smart default (%)	Direct prediction (%)	Privacy attitude (%)	Social behavior (%)	Negotiability (%)
G set						
SN public	90.00	90.00	90.00	90.00	90.00	90.00
SN friends Only	73.33	53.33	73.33	63.33	60.00	56.67
Health	66.67	60.00	60.00	43.33	40.00	70.00
Other apps	76.67	76.67	76.67	76.67	76.67	76.67
Corporate	80.00	80.00	80.00	80.00	80.00	80.00
Government	86.67	86.67	86.67	86.67	86.67	86.67
Health	86.67	86.67	86.67	86.67	86.67	86.67
Safety	90.00	90.00	90.00	90.00	90.00	90.00
Social	93.33	60.00	100.00	70.00	60.00	63.33
Commercial	73.33	73.33	73.33	73.33	73.33	73.33
Convenience	80.00	73.33	73.33	76.67	66.67	70.00
Frequency	53.33	53.33	53.33	53.00	53.33	53.33
Retention	50.00	40.00	50.00	50.00	43.33	46.67
Average	76.92	71.02	76.41	72.31	69.74	72.56
Overall average	84.74	68.74	83.41	68.52	69.70	73.11

References

Abhigna, B., Soni, N., Dixit, S.: Crowdsourcing—a step towards advanced machine learning. Proc. Comput. Sci. **132**, 632–642 (2018)

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

Agarwal, Y., Hall, M.: 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, pp. 97–110. ACM (2013)

Almuhimedi, H., Schaub, F., Sadeh, N., Adjerid, I., Acquisti, A., Gluck, J., Cranor, L.F., Agarwal, Y.: Your location has been shared 5398 times!: A field study on mobile app privacy nudging. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, pp. 787–796. ACM (2015)

Assad, M., Carmichael, D., Kay, J., Kummerfeld, B.: Giving users control over location privacy. In: Workshop on Ubicomp Privacy (2007)

Bahirat, P., He, Y., Menon, A., Knijnenburg, B.: A data-driven approach to developing iot privacy-setting interfaces. In: 23rd International Conference on Intelligent User Interfaces, pp. 165–176. ACM (2018)

Bellotti, V., Sellen, A.: Design for privacy in ubiquitous computing environments. In: Proceedings of the Third European Conference on Computer-Supported Cooperative Work, 13–17 September 1993, Milan, Italy ECSCW'93, pp. 77–92. Springer (1993)

Beresford, A.R., Rice, A., Skehin, N., Sohan, R.: Mockdroid: trading privacy for application functionality on smartphones. In: Proceedings of the 12th Workshop on Mobile Computing Systems and Applications, pp. 49–54. ACM (2011)

Brank, J., Grobelnik, M., Mladenić, D.: A survey of ontology evaluation techniques. In Proceedings of the conference on data mining and data warehouses (SiKDD 2005). Ljubljana, Slovenia, pp. 166–170 (2005)

Brar, A., Kay, J.: Privacy and Security in Ubiquitous Personalized Applications. University of Sydney, School of Information Technologies, Sydney (2004)

Carmagnola, F., Osborne, F., Torre, I.: Escaping the big brother: an empirical study on factors influencing identification and information leakage on the web. J. Inf. Sci. **40**(2), 180–197 (2014)

Chakraborty, S., Shen, C., Raghavan, K.R., Shoukry, Y., Millar, M., Srivastava, M.B.: ipshield: A framework for enforcing context-aware privacy. In: NSDI, pp. 143–156 (2014)

Chaturvedi, A., Green, P.E., Caroll, J.D.: K-modes clustering. J. Classif. **18**(1), 35–55 (2001)

Chaudhry, A., Crowcroft, J., Howard, H., Madhavapeddy, A., Mortier, R., Haddadi, H., McAuley, D.: Personal data: thinking inside the box. In: Proceedings of The Fifth Decennial Aarhus Conference on Critical Alternatives, pp. 29–32. Aarhus University Press (2015)

Conger, S., Pratt, J.H., Loch, K.D.: Personal information privacy and emerging technologies. Inf. Syst. J. **23**(5), 401–417 (2013). https://doi.org/10.1111/j.1365-2575.2012.00402.x

Dinev, T., Hart, P.: An extended privacy calculus model for e-commerce transactions. Inf. Syst. Res. **17**(1), 61–80 (2006)

Egele, M., Kruegel, C., Kirda, E., Vigna, G.: Pios: Detecting privacy leaks in ios applications. In: NDSS, pp. 177–183 (2011)

Elluri, L., Joshi, K.P., et al.: A knowledge representation of cloud data controls for EU GDPR compliance. In: 11th IEEE International Conference on Cloud Computing (CLOUD) (2018)

Felt, A.P., Ha, E., Egelman, S., Haney, A., Chin, E., Wagner, D.: Android permissions: user attention, comprehension, and behavior. In: Proceedings of the Eighth Symposium on Usable Privacy and Security, pp. 1–14. ACM (2012)

Fu, H., Yang, Y., Shingte, N., Lindqvist, J., Gruteser, M.: A field study of run-time location access disclosures on android smartphones. Proc. Usable Secur. **14**, 10 (2014)

Google/Ipsos, U.: How people discover, use, and stay engaged with apps, pp. 1–15 (2016). https://www.thinkwithgoogle.com/data/smartphone-users-discover-apps-browsing/

He, Y., Bahirat, P., Menon, A., Knijnenburg, B.P.: A data driven approach to designing for privacy in household iot. ACM Trans. Interact. Intell. Syst. **10**(1) (2019)

Hlomani, H., Stacey, D.: Approaches, methods, metrics, measures, and subjectivity in ontology evaluation: a survey. Semant. Web J. **1**(5), 1–11 (2014)

Johnson, E.J., Bellman, S., Lohse, G.L.: Defaults, framing and privacy: why opting in-opting out. Market. Lett. **13**(1), 5–15 (2002)

Joosse, M., Lohse, M., Evers, V.: Crowdsourcing culture in HRI: Lessons learned from quantitative and qualitative data collections. In: 3rd International Workshop on Culture Aware Robotics at ICSR, vol. 15 (2015)

Kay, J., Kummerfeld, B.: Scrutability, user control and privacy for distributed personalization. In: Proceedings of the CHI2006 Workshop on Privacy-Enhanced Personalization, pp. 21–22 (2006)

Kay, J., Kummerfeld, B., Lauder, P.: Personis: a server for user models. In: International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems, pp. 203–212. Springer (2002)

Kelley, P., Consolvo, S., Cranor, L., Jung, J., Sadeh, N., Wetherall, D.: A conundrum of permissions: installing applications on an android smartphone. In: International conference on Financial Cryptography and Data Security, Springer, Berlin, Heidelberg, pp. 68–79 (2012)

Knijnenburg, B., Raybourn, E., Cherry, D., Wilkinson, D., Sivakumar, S., Sloan, H.: Death to the privacy calculus? (2017). Available at SSRN:http://dx.doi.org/10.2139/ssrn.2923806

Knijnenburg, B.P.: Information disclosure profiles for segmentation and recommendation. In: SOUPS2014 Workshop on Privacy Personas and Segmentation (2014)

Knijnenburg, B.P.: A user-tailored approach to privacy decision support. Ph.D. Thesis, University of California, Irvine (2015). http://search.proquest.com/docview/1725139739/abstract

Knijnenburg, B.P.: Privacy? I can't even! Making a case for user-tailored privacy. IEEE Secur. Privacy **15**(4), 62–67 (2017)

Knijnenburg, B.P., Jin, H.: The persuasive effect of privacy recommendations. In: Twelth Annual Workshop on HCI Research in MIS, Milan (2013). http://aisel.aisnet.org/sighci2013/16

Knijnenburg, B.P., Kobsa, A.: Helping users with information disclosure decisions: potential for adaptation. In: Proceedings of the 2013 International Conference on Intelligent User Interfaces, pp. 407–416. ACM (2013)

Knijnenburg, B.P., Kobsa, A., Jin, H.: Counteracting the negative effect of form auto-completion on the privacy calculus. In: ICIS 2013 Proceedings, Milan (2013)

Knijnenburg, B.P., Kobsa, A., Jin, H.: Dimensionality of information disclosure behavior. Int. J. Hum. Comput. Stud. **71**(12), 1144–1162 (2013). https://doi.org/10.1016/j.ijhcs.2013.06.003

Kobsa, A.: Tailoring privacy to users' needs. In: International Conference on User Modeling, pp. 301–313. Springer (2001)

Kodinariya, T.M., Makwana, P.R.: Review on determining number of cluster in k-means clustering. Int. J. **1**(6), 90–95 (2013)

Kurtz, C., Semmann, M., Schulz, W.: Towards a framework for information privacy in complex service ecosystems. In: ICIS 2018 Proceedings (2018). https://aisel.aisnet.org/icis2018/bridging/Presentations/7

Lee, H., Kobsa, A.: Understanding user privacy in internet of things environments. In: 2016 IEEE 3rd World Forum on Internet of Things (WF-IoT), pp. 407–412. IEEE (2016)

Lee, H., Kobsa, A.: Privacy preference modeling and prediction in a simulated campuswide iot environment. In: IEEE International Conference on Pervasive Computing and Communications (PerCom), pp. 276–285. IEEE (2017)

Li, Y., Kobsa, A., Knijnenburg, B.P., Nguyen, M.C.: Cross-cultural privacy prediction. Proc. Privacy Enhanc. Technol. **2017**(2), 113–132 (2017)

Lin, J., Liu, B., Sadeh, N., Hong, J.I.: Modeling users' mobile app privacy preferences: restoring usability in a sea of permission settings. In proceedings of the 10th Symposium On Usable Privacy and Security (SOUPS), pp. 199–212 (2014)

Liu, B., Andersen, M.S., Schaub, F., Almuhimedi, H., Zhang, S., Sadeh, N., Acquisti, A., Agarwal, Y.: Follow my recommendations: A personalized privacy assistant for mobile app permissions. In: Twelfth Symposium on Usable Privacy and Security, pp. 26–41 (2016)

Liu, B., Lin, J., Sadeh, N.: Reconciling mobile app privacy and usability on smartphones: Could user privacy profiles help? In: Proceedings of the 23rd International Conference on World Wide Web, pp. 201–212. ACM (2014b)

Madejski, M., Johnson, M., Bellovin, S.: A study of privacy settings errors in an online social network. In: Fourth International Workshop on Security and Social Networking, SECSOC '12, pp. 340–345. Lugano (2012). https://doi.org/10.1109/PerComW.2012.6197507

Malhotra, N.K., Kim, S.S., Agarwal, J.: Internet users' information privacy concerns (IUIPC): the construct, the scale, and a causal model. Inf. Syst. Res. **15**(4), 336–355 (2004)

Noy, N.F., McGuinness, D.L., et al.: Ontology development 101: a guide to creating your first ontology. Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880, March 2001 (2001) http://www.ksl.stanford.edu/people/dlm/papers/ontology-tutorial-noy-mcguinness.pdf

Palmirani, M., Martoni, M., Rossi, A., Bartolini, C., Robaldo, L.: Pronto: privacy ontology for legal reasoning. In: International Conference on Electronic Government and the Information Systems Perspective, pp. 139–152. Springer (2018)

Pandit, H., Lewis, D.: Modelling provenance for GDPR compliance using linked open data vocabularies. In: 5th Workshop on Society, Privacy and the Semantic Web - Policy and Technology (PrivOn 2017), CEUR 1951 (2017). http://ceur-ws.org/Vol-1951/#paper-06

Pandit, H.J., Fatema, K., O'Sullivan, D., Lewis, D.: Gdprtext-gdpr as a linked data resource. In: European Semantic Web Conference, pp. 481–495. Springer (2018)

Patil, T.R., Sherekar, S.: Performance analysis of naive bayes and j48 classification algorithm for data classification. Int. J. Comput. Sci. Appl. **6**(2), 256–261 (2013)

Perera, C., Liu, C., Ranjan, R., Wang, L., Zomaya, A.Y.: Privacy-knowledge modeling for the internet of things: a look back. Computer **49**(12), 60–68 (2016)

Raber, F., Krüger, A.: Deriving privacy settings for location sharing: Are context factors always the best choice? In: 2018 IEEE Symposium on Privacy-Aware Computing (PAC), pp. 86–94. IEEE (2018)

Rafailidis, D., Nanopoulos, A.: Modeling users preference dynamics and side information in recommender systems. IEEE Trans. Syst. Man Cybern. Syst. **46**(6), 782–792 (2016)

Sacco, O., Breslin, J.G.: Ppo & ppm 2.0: extending the privacy preference framework to provide finer-grained access control for the web of data. In: Proceedings of the 8th International Conference on Semantic Systems, pp. 80–87 (2012)

Sanchez, O., Torre, I., Knijnenburg, B.: Semantic-based privacy settings negotiation and management. In: Future Generation Computer Systems (2019). (Under publication)

Schemmann, B., Herrmann, A.M., Chappin, M.M., Heimeriks, G.J.: Crowdsourcing ideas: involving ordinary users in the ideation phase of new product development. Res. Policy **45**(6), 1145–1154 (2016)

Sharma, S., Chen, K., Sheth, A.: Toward practical privacy-preserving analytics for iot and cloud-based healthcare systems. IEEE Internet Comput. **22**(2), 42–51 (2018)

Si, C., Jiao, L., Wu, J., Zhao, J.: A group evolving-based framework with perturbations for link prediction. Physica A **475**, 117–128 (2017)

Smith, H.J., Milberg, S.J., Burke, S.J.: Information privacy: measuring individuals' concerns about organizational practices. MIS Quarterly: Management Information Systems **20**(2), 167–196 (1996)

Sutanto, J., Palme, E., Tan, C.H., Phang, C.W.: Addressing the personalization-privacy paradox: an empirical assessment from a field experiment on smartphone users. Mis Quart. **37**(4), 1141–1164 (2013)

The European Parliament and the Council of the European Union: Regulation (EU) 2016/679 of the European Parliament and of the Council. Official Journal of the European Union, p. 1:88 (2016)

Torre, I., Adorni, G., Koceva, F., Sanchez, O.: Preventing disclosure of personal data in iot networks. In: 12th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), pp. 389–396. IEEE (2016a)

Torre, I., Koceva, F., Sanchez, O.R., Adorni, G.: Fitness trackers and wearable devices: How to prevent inference risks? In: Proceedings of the 11th EAI International Conference on Body Area Networks, pp. 125–131. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering) (2016b)

Torre, I., Koceva, F., Sanchez, O.R., Adorni, G.: A framework for personal data protection in the iot. In: 11th International Conference for Internet Technology and Secured Transactions (ICITST), pp. 384–391. IEEE (2016c)

Torre, I., Sanchez, O.R., Koceva, F., Adorni, G.: Supporting users to take informed decisions on privacy settings of personal devices. Pers. Ubiquit. Comput. **22**(2), 345–364 (2018)

Tsai, L., Wijesekera, P., Reardon, J., Reyes, I., Egelman, S., Wagner, D., Good, N., Chen, J.W.: Turtle guard: helping android users apply contextual privacy preferences. In: Symposium on Usable Privacy and Security (SOUPS) (2017)

Vescovi, M., Moiso, C., Pasolli, M., Cordin, L., Antonelli, F.: Building an eco-system of trusted services via user control and transparency on personal data. In: IFIP International Conference on Trust Management, pp. 240–250. Springer (2015)

Vicente, C.R., Freni, D., Bettini, C., Jensen, C.S.: Location-related privacy in geo-social networks. IEEE Internet Comput. **15**(3), 20–27 (2011)

Walters, M.L., Lohse, M., Hanheide, M., Wrede, B., Syrdal, D.S., Koay, K.L., Green, A., Hüttenrauch, H., Dautenhahn, K., Sagerer, G., et al.: Evaluating the robot personality and verbal behavior of domestic robots using video-based studies. Adv. Robot. **25**(18), 2233–2254 (2011)

Wijesekera, P., Baokar, A., Tsai, L., Reardon, J., Egelman, S., Wagner, D., Beznosov, K.: The feasibility of dynamically granted permissions: aligning mobile privacy with user preferences. In: IEEE Symposium on Security and Privacy (SP), pp. 1077–1093. IEEE (2017)

Wisniewski, P., Knijnenburg, B.P., Lipford, H.R.: Profiling facebook users privacy behaviors. In: SOUPS2014 Workshop on Privacy Personas and Segmentation (2014)

Woods, S., Walters, M., Koay, K.L., Dautenhahn, K.: Comparing human robot interaction scenarios using live and video based methods: towards a novel methodological approach. In: 9th IEEE International Workshop on Advanced Motion Control, pp. 750–755. IEEE (2006)

Wu, L., Ge, Y., Liu, Q., Chen, E., Hong, R., Du, J., Wang, M.: Modeling the evolution of users' preferences and social links in social networking services. IEEE Trans. Knowl. Data Eng. **29**(6), 1240–1253 (2017)

Wu, L., Ge, Y., Liu, Q., Chen, E., Long, B., Huang, Z.: Modeling users' preferences and social links in social networking services: a joint-evolving perspective. In: Thirtieth AAAI Conference on Artificial Intelligence (2016)

Xie, J., Knijnenburg, B.P., Jin, H.: Location sharing privacy preference: analysis and personalized recommendation. In: Proceedings of the 19th international conference on Intelligent User Interfaces, pp. 189–198. ACM (2014)

Xu, H., Dinev, T., Smith, H.J., Hart, P.: Examining the formation of individual's privacy concerns: toward an integrative view. In: ICIS 2008 Proceedings, p. 6 (2008)

Xu, H., Gupta, S., Rosson, M.B., Carroll, J.M.: Measuring mobile users' concerns for information privacy, Proc. of the Third International Conference on Information Systems, Orlando, pp. 2278–2293 (2012)

Zhao, Y., Zhu, Q.: Evaluation on crowdsourcing research: current status and future direction. Inf. Syst. Front. **16**(3), 417–434 (2014)

Zhao, Z., Etemad, S.A., Arya, A.: Gamification of exercise and fitness using wearable activity trackers. In: Proceedings of the 10th International Symposium on Computer Science in Sports (ISCSS), pp. 233–240. Springer (2016)

Odnan Ref Sanchez is a current postdoctoral fellow in the University of Strasbourg. He obtained his PhD in Digital Humanities and master's degree in Telecommunication's Engineering at the University of Genoa, Italy. He obtained his bachelor's degree in Electronics and Communication's Engineering at CIT-University in Cebu City, Philippines. His previous research includes privacy recommendation in IoT, user and network modeling, and green NFV.

Ilaria Torre is an Assistant Professor at the Department of Informatics, Bioengineering, Robotics, and Systems Engineering, University of Genoa, and a member of the Advisory Board of the PhD School in Digital Humanities. Her main research activity is in the field of intelligent and adaptive user interfaces, applied to different domains including on-board services and tourist–cultural mobile guides, technology-enhanced learning, and privacy management. Over the years, she served as a TPC member of several conferences and as co-organizer of international events. She is general chair of ACM UMAP 2020 (User Modeling, Adaptation, and Personalization).

Yangyang He is currently a PhD Candidate in Computer Science at Clemson University. He received a B.E. in Computer Engineering from Beihang University, China, and an M.S. in Computer Science from the Clemson University. Yangyang works on utilizing machine learning algorithms to help IoT users better manage their privacy.

Bart Knijnenburg is an Assistant Professor in Human-Centered Computing at the Clemson University School of Computing where he co-directs the Humans and Technology laboratory. He holds a B.S. in Innovation Sciences and an M.S. in Human–Technology Interaction from the Eindhoven University of Technology, The Netherlands, an M.A. in Human–Computer Interaction from Carnegie Mellon University, and a PhD in Information and Computer Sciences from UC Irvine. Bart works on privacy decision-making and user-centric evaluation of adaptive systems. His research has received funding from the National Science Foundation, the Department of Defense, and corporate sponsors.

Affiliations

Odnan Ref Sanchez[1] · **Ilaria Torre**[1] · **Yangyang He**[2] · **Bart P. Knijnenburg**[2]

Odnan Ref Sanchez
odnan.ref.sanchez@edu.unige.it

Yangyang He
yyhe@g.clemson.edu

Bart P. Knijnenburg
bartk@clemson.edu

[1] Department of Computer Science, Bioengineering, Robotics and Systems Engineering (DIBRIS), University of Genoa, Genoa, Italy

[2] School of Computing, Clemson University, Clemson, USA