UNIVERSITÀ DEGLI STUDI DI GENOVA

DOCTORAL THESIS

A Framework to Support Users' Privacy Preferences in the Proliferation of IoT

Author:
Odnan Ref SANCHEZ

Supervisor:
Ilaria TORRE

A thesis submitted in fulfillment of the requirements
for the degree of Doctor of Philosophy

in

Digital Humanities (31st cycle)

Artificial Intelligence and Multi-agent Systems (AIMS) Laboratory
Department of Informatics, Bioengineering, Robotics, and Systems Engineering (DIBRIS)

May 29, 2019

"He who laughs, lasts."

Odnan Ref Sanchez

UNIVERSITÀ DEGLI STUDI DI GENOVA

Abstract

A Framework to Support Users' Privacy Preferences in the Proliferation of IoT

by Odnan Ref SANCHEZ

In the proliferation of personal IoT devices, the need for privacy protection becomes an increasing concern. User's privacy preferences are not being respected in today's complex IoT scenario, as data sharing among applications becomes a growing phenomenon. The increasing number of applications, IoT devices and list of user's personal data make the setting of privacy a laborious task for the users. On the other hand, supposedly trusted third parties that access personal data have been recently reported to invade user privacy. Thus, this thesis proposes a privacy framework that computes the risk of users' sharing preferences, manages user privacy and provides recommendation to ease privacy setting in the advent of IoT. The risk of inferencing unshared user data is computed from the set of shared user data. The framework aims to be GDPR-compliant, which makes third parties declare their access request in accordance with the European Union's General Data Protection Regulation (GDPR). Semantic Web Technologies are used to model both the user and the third party preferences, which can be represented through the proposed Privacy Preference for the IoT (PPIoT) Ontology. The framework's personal data manager supports the privacy decision of the user through recommendation of privacy profiles. Using Machine Learning techniques, the identification and recommendation of privacy profiles are done through our crowdsourced dataset, which are collected using current scenarios in the fitness domain. We then examine different personal tracking data and user traits which can potentially drive the recommendation of privacy profiles to the users. Interestingly, our results show several semantic relationships among users' traits, characteristics and attitudes that are useful in providing privacy recommendations.

Acknowledgements

I am forever grateful to the University of Genoa for all the opportunities that have been given to me. Thanks for selecting me as a scholar both during the master and PhD program. Throughout my PhD program, I am thankful to have a great mentor, Prof. Ilaria Torre, who have supported me all this time and have created a lot of opportunities in order for me to grow. I will always be indebted to you and will always be grateful to be your student. I also thank Prof. Franco Davoli for all the unwavering support since the master program days and Prof. Giovanni Adorni, our PhD coordinator, for the support and guidance during the PhD.

A special thanks to Prof. Bart P. Knijnenburg for making my PhD abroad possible and for sharing your expertise with us. I am proud to be part of your team for a short while.

To my family, especially my mother, Evelyn D. Sanchez, I dedicate this work to you. Thank you for always inspiring me. Finally, I thank my ever-supportive wife-to-be, Dr. Jane Frances "lablab" Pajo, for being there for me always.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Today, Internet of Things (IoT) has been steadily increasing and has been widely deployed. These things and sensors, which have the capability to connect to the network, undoubtedly aid people in their everyday activities. IoT is transforming and helping people's way of life, improving traditional businesses and facilitating the progress of smart cities. Though the advantages are clear, one of the main concerns toward IoT is the users' privacy. Having an unprecedented ability to sense, control, collect and process users' personal data, IoT poses a lot of risks on user privacy.

Preserving the privacy of the users in the context of the IoT is a challenging task. 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.

This thesis aims to bridge the gap concerning the current privacy issues in the IoT paradigm by proposing a framework that manages user privacy in IoT, providing ease of access and configuration of privacy settings. The framework acts as an intermediary between the user and the third parties, managing the users' privacy preferences. Additionally, the framework also provides information on the potential inferences of undisclosed data, given the set of personal data that users have disclosed to third parties.

The framework also takes into account the newly adopted EU General Data Protection Regulation (GDPR) [61], which represents the most important change in data privacy laws in the last twenty years. While the GDPR is a significant stride towards

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*.

In the following sections, the scope and definition of privacy will be demonstrated. This provides insights on the domain of study and narrows down to the specific privacy problems that this study aims to focus. Then, the potential privacy concerns in selected Digital Humanities domains will be introduced, which show the importance of privacy in different areas due to the IoT proliferation. Then, a brief introduction of the framework will be discussed, which introduces the main research question of this thesis. Finally, the dissertation organization will be presented.

1.1 Privacy Definition

Privacy has been long studied in literature. A well-known and well-used definition of privacy can be found from Prof. Westin [193], which states that privacy is the right of the claim of individuals, groups, or institutions to determine for themselves when, how, and to what extent information about them is communicated to others. Therefore, privacy highly depends on the perception of the entity, which decides if certain data is considered private or not.

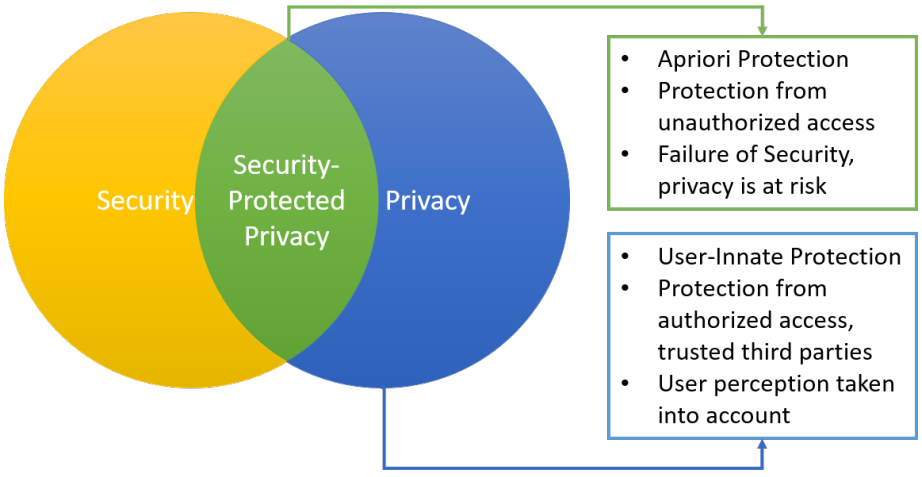


FIGURE 1.1: The conceptual intersection between security and privacy.

IoT is expected to introduce numerous challenges especially privacy and security issues [90]. In literature, privacy protection is mostly coupled in the security domain. However, security methods only protect parts of the whole user privacy domain as shown in Figure 1.1. To clarify the relation between security and privacy, a basic overview of the definition is provided in the figure. The overlap of security and privacy (green) is defined as security-protected privacy, which has the aim of protecting user's privacy without taking into account the user's individual privacy perception. It provides a priori protection ensuring the general user privacy. It is usually secured through the authentication and authorization mechanisms that aims in protecting the general user data, which may be private for a given user. When security protection fails, user privacy is at risk. But when security is successful, it does not mean that complete user privacy is not at risk any more. Privacy of users must still be protected continually given that a number of recent privacy breaches are due to the allowed and supposedly trusted access of third parties (e.g., [106]). More importantly, individual user's privacy perception must also be considered (blue part in Figure 1.1), following Westin's definition.

Privacy is considered as one of the important security principles, making the overlap in Figure 1.1 arguably bigger, due to a large number of devices, services, and people sharing the same communication network in IoT [116]. Privacy concerns increase in the introduction of IoT, mainly because these devices can collectively gather massive amounts of personal information, without properly informing users or even asking for their permission [126, 37, 189, 113]. This security-related privacy domain has numerous studies in literature.

In conclusion, this thesis will focus on privacy by Westin's definition and will not take into account any security-related privacy matter. The goal of the privacy protection part of the framework is to protect user's from potential inference of undisclosed data, which are not permitted by the users and can be derived from the data shared to the third parties. This issue becomes alarming nowadays since a number of supposedly trusted third parties have been reported to violate user privacy (e.g., [181]). Additionally, a part of the framework provides privacy recommendation to support the setting of user privacy in the complex IoT scenario.

1.2 Privacy in Digital Humanities

The technological shift introduced by the IoT in different realms of discipline, such as in Digital Humanities, needs further attention to privacy. In this section, selected domains of Digital Humanities which have potential privacy concerns are presented. As IoT sensors and applications are mostly used in Digital Humanities, it is therefore also subject to privacy concerns in the digital age. In this section, we briefly discuss how the introduction of new sensors will affect Digital Humanities domains.

1.2.1 Digital Tourism

Travelling for leisure or business purposes can be exhaustive. To this point, digital tourism aims in improving the quality of the traveller experience by providing them digital support [185]. These solutions and approaches are in form of recommender systems, tools or apps. Many of the solution use GPS-based systems (e.g., interactive maps, tourism assistants) and augmented reality (i.e., new method of presenting historical content for tourist locations) using the position and orientation of sensors on smartphones [16, 159].

IoT enables smart tourism by sensing geographical things, natural disasters, tourist behaviors, infrastructure of the scenic spots, etc. [69] This is possible through the applications of RFID technology, sensor network, mobile systems, supporting technologies of IoT in socio-economic life, and IoT tracking systems, which could identify real-time positioning of visitors. Among others, tourist destination selection, routes planning optimization, hotel and travel bookings, and integration management of tourist attractions can be included in the IoT smart tourism [69].

Obviously, concerns about privacy rise from the persistent access of user data. Concerns multiplied as real-time location-based services collect significant amounts of real-time information in order to deliver the app functions needed by the users [103]. As stated in Gretzel et al. [67], data lies at the core of all smart tourism activities. The need for information is so great in digital tourism that tourists might be easily persuaded to forfeit privacy. Therefore, information governance and privacy are major areas of research necessary in the context of digital tourism [67].

1.2.2 Assisted Living

Ambient assisted living (AAL) is based on ambient intelligence that are sensitive to human environments [151]. AAL is widely used for monitoring, curing, and improving wellness and health conditions of the elderly. Additionally, AAL provides safety and security through emergency response systems and fall detections systems. AAL also provides better connectivity among family and friends of the elderly. IoT systems enabling AAL mostly composed of body sensors (e.g., heartrate monitors, temperature, blood pressure, sweat senors), home monitoring systems (e.g, CO_2, temperature sensors), light and video surveillance systems, medication control systems, and mobile systems that are persistently monitoring the elderly.

While the advantages are clear, the intensive monitoring of the elderly also imposes risks on privacy. Medical data are already privacy sensitive and the introduction of high precision real-time access to provide the needed services for elderly makes it even more concerning.

This concern has been studied in Wood et al. [201], providing privacy management features for AAL. The ALARM-NET is built for wireless sensors in AAL that could monitor the environmental and physiological data of the elderly. For the privacy, it incorporates a Circadian Activity Rhythm (CAR) analysis module that learns the patterns of daily life of the elderly. This enables to dynamically configure privacy settings for the elderly, which is triggered when they exhibit a behavior that is critical to their health and enable the authorized medical personnel to access vital data. The data can be hidden or available for anonymous statistical purposes as managed by the privacy manager.

1.2.3 Digital Museum

Museums are also being transformed digitally to accommodate the needs of the intended users [125], collecting digital resources at a rapid pace given the inadequacy of user data. Understanding the user needs has been one of the goals of the digital museums to provide satisfactory user experience. Providing a more user-centric online resource has been deemed important among museum professionals which must

be understood from online visitors and how they are using such resources for their activities.

An example of a personalized museum is studied in Koshizuka et al. [102]. The authors state that at least the differences of general adults, experts and school children must be taken into account. Additionally, the disabilities such as color blindness, weak eyesight, physical conditions, must also be considered to provide a satisfiable experience to different visitors. Personalized services can be provided using the visitor's data, which can be digitally transformed through the RFID technology [102]. RFID readers are provided in front of the exhibits and visitors can put their personal RFID card into the system which connects and queries the network to provide personalized services.

According to Hsi et al. [77], the usage of RFID technologies for smart museums will likely expand, as it provides enormous promise for improved visitor learning. Museums, however, must educate their visitors about the RFID and manage the associated privacy risks, in addition to developing interactive exhibition designs. Educating users across all fields is needed to help users improve their knowledge on sharing their private information and the associated risks. Furthermore, there are also contradictory messages coming from the visitors. Survey shows that museum visitors say they want more privacy, yet when they are asked to opt-in, they are willing [82].

In addition to RFID system, Karaman et al. [84] also used camera network and mobile system to persistently monitor each visitor and provide personalized multimedia content. For each visitor, the system estimates a profile vector describing the visitor's appearance for re-identification purposes and estimate his interest profile. Not only does the persistency of data access impose user privacy, but also the level of detail on the user's profile. Thus, privacy in digital museums is becoming a concern in the proliferation of IoT.

On the security-protected privacy point of view, museums also have difficulties in ensuring the privacy of newly acquired data from external attacks [82]. They emphasize their favour on museum programs that accumulate user data, encrypt credit card number, and unlink them with their user's profile. However, these assurances

do not guarantee from cyber attacks. Digital museums are also being targeted nowadays such as the recent (2014) case on Ashmolean Museum, a university museum on the Oxford campus. Protecting personal information is really a challenge as it is the currency of the knowledge economy and the key focus of today's several business models.

1.2.4 Digital Hospitals

Digital hospitals provide a more efficient medication service to patients in hospitals. Closed-loop Electronic Medication Management Systems (EMMS) provides electronic prescribing, automated dispensing, barcode patient identification and the use of an electronic medication administration record (eMAR) [9]. It improves efficiency in medication, which reduces errors in prescription from 3.8% to 2.0% and medication administration errors from 8.6% to 4.4%. In a critical situation like hospitals, efficiency is obviously necessary.

While having innovations in hospital health systems, strong person-centered approach are still maintained as it is important to understand individual patient's need [26]. Digital records should represent an accurate patient information. In digital hospital scenario, technology is enabling medical health records to be put in the electronic format, EPRs, and making them accessible by the users through the network. Additionally, with the help of IoT sensor networks, the idea of remote patient monitoring is now possible [127]. Examples of persistent monitors that connect to the network in the domain of Internet of Medical Things (IoMT) [212] include thermister (temperature), lethysmograph (respiratory monitor), scale (weight), motion sensors (movement), blood sensor, galvanic response (anxiety detector), skin conductance (sweat sensor), muscle contraction, etc. In this scenario, it is clear that detailed and sensitive personal data types are collected dynamically which make privacy a growing concern.

Digital medical records are advantageous, as it gives ease in efficiency and ease of processing. However, the benefits of the technology must also trade-off the privacy of the patients. Current privacy issues also include access rights to data, how and when data is stored, security of data transfer, data analysis rights, and the governing

policies [127]. While medical data are under regulation, it must be re-evaluated periodically as technology improves rapidly, which changes how health care systems are being implemented.

1.2.5 Online Learning

Online learning uses the technology to bring together remote learners all around the world. Distributed learners interact and learn through platforms and mingle with the help of sensor environments. Today, it becomes popular in handling massive user data as its advantage becomes evident [208].

Online learning evolved in to mobile learning (m-learning) where learners have the freedom to move, and recently shifted in to context-aware distant learning in which the system can detect the student's learning behaviours through IoT [119]. This enables a more effective and personalized learning by guiding them to the right target areas, provide materials and evaluate their performance. Students access through their computer, mobile devices, personal digital assistants (PDAs), sensor technologies (e.g., RFID readers, tags, GPS). Additionally, brain monitors such as Electroencephalogram (EEG) can be used to measure the learner's attention level [112] in real-time. EEG is known for attention recognition using brain wave patterns. Furthermore, heart rate monitor, skin conductance monitor, blood volume pressure monitor, also allow to monitor user's emotion information which is used for a more effective learning [164]. A positive mood makes the learner more creative and flexible in problem solving, and more efficient in decision making.

Online learning has become more efficient due to the proliferation of IoT. However, it also introduce new personal information that are sensitive. Thus, privacy of learners are also at risk, with sensor systems and platforms constantly monitoring them during the learning activity.

1.3 Main Goal and Solution

Given today's privacy urgency as discussed above, privacy models have also been widely studied. Privacy can be managed by means of a data manager which allows the user, or a trusted entity to control the personal data and the policies to make

them available to third parties, as proposed in previous studies (e.g., [33]). This requires agreements and standardized procedures since individual providers and their applications have to interface with a data manager. Several frameworks for managing user data have been proposed in the last years [33, 218, 187].

However, regulating the access of one's data cannot deny the possible disclosure through inference. Management does not assure that privacy will be fully respected by the third parties. Privacy is not simply about encryption and access authorization [40], but also about breaking data correlation so that the risk of inference is reduced. Inference attacks are based on the integration and correlation of known data about an individual who leads to the discovery of private data by the supposedly trusted third parties. The inference is done by linking sensitive information to the knowledge that may be available to a third party as a background knowledge, common sense or domain-specific knowledge [149].

In this thesis, the fundamental research question that is investigated is as follows:

> *How can the privacy of the users be easily managed in the complex IoT scenario and how can we prevent accessing third parties that are permitted by the user to prevent further data access?*

To answer this question, this thesis focuses on creating a framework that could manage user privacy settings and recommend privacy profiles that are best-suited to each user. These settings aims to be GDPR-compliant following the main principles for personal data handling. Then, the framework will also be able to compute risks associated to sharing sets of user's personal data by computing the correlation of these data to the unshared data.

In this framework, the protection of user privacy has two main functions. The privacy framework therefore divides the capability in two modules, namely, the Personal Data Manager (PDM) and the Adaptive Inference Discovery System (AID-S). The PDM manages user privacy in the complex IoT scenario while the AID-S provides measures regarding the risks of data inference.

Torre et al. [181] points out the excessive access of personal data in the fitness IoT. Currently, among the Digital Humanities domains, digital fitness is becoming popular as millions of users engage in this domain. Also, multiple personal data

are being accessed persistently on round-the-clock activities by fitness apps. For this reason, the framework will be applied to the digital fitness domain in this study.

1.4 Dissertation Organization

This dissertation is organized as follows. The next chapter provides related work in privacy in the IoT scenario that are relevant to this thesis. Chapter 3 gives the full details of the PerNANDO Framework, which is the proposed privacy framework for IoT. Chapter 4 discusses the Fitness IoT use case, which is the domain of IoT chosen for this study. In Chapter 5, the AID-S component of the framework will be discussed, focusing on the inference risks of sharing personal data. The PDM framework will be discussed in Chapters 6 and 7. In Chapter 6, the PPIoT Ontology is defined and how it can be used by the PDM to model privacy preferences. Then, the PDM data modeling, privacy profiling and recommendation strategies are elaborated in Chapter 7. Finally, the concluding remarks, limitations, future work, and open issues are discussed in Chapter 8.

Chapter 2

Background and Related Work

In this chapter, the discussion is focused on two main parts, namely, inference attacks, and privacy management. The first part briefly discussed the current works on inference protection, its measures and categories while the succeeding part discuss the privacy management, privacy preference modeling and recommendation.

2.1 Inference Attacks

Inference attacks are usually described as private information disclosure based on the integration and correlation of non-private data [5]. They are also described as a form of intense data mining where confidential information is harvested and disclosed by integrating nonsensitive data with metadata.

This issue is not new since it has been studied in OSNs [5, 73, 34, 28, 29] and in database management [43, 176, 179]. The IoT and the Web 3.0 increase the risks of inference attacks since there is a massive increase of personal data distributed on the network. Also, the growth of richer metadata increases the possibility of data integration and correlation. Users may lose track of sensitive data available on the network and where it is stored, which may lead to an increase in the precision and volume of inference attacks [25]. Finally, users may not be aware of the digital footprints left on different service providers [191]. The risk imposed is through the combination of user data from different sources and infer new user information.

Ahmadinejad et al. [5] reveal that success rates of inference algorithms in systems that share user data were high, which make it alarming in the personal IoT context. It was also shown that inference attacks can be employed as a building block for

launching further security attacks, including, for example, phishing and bypassing authentication challenges.

Below, we discuss the main inferences derived from authorized/unauthorized acquisition of data, the common measures of inferences, and the current methods to prevent inferences.

2.1.1 Inference attacks from unauthorized acquisition

Many studies on privacy protection are about protecting users from external attackers (i.e., external inference attacks). In mobile systems, external inference attacks are prevented from the vulnerability of the operating systems. For example, several studies analyze the weakness of Android operating system which potentially infer the user's sensitive information (e.g., [202, 54, 74, 130, 10, 215]) . Android's openness makes users susceptible to privacy risks. Given this weakness, this group of studies aims to defend users mainly from external attacks and malware penetration.

Other than inference on mobile systems, studies about external inference attacks in the IoT show how the vulnerabilities of wireless networks can be exploited to discover private data (e.g., [68]). A typical example is acquiring data on electricity consumption in a building to infer the user habits. For instance, Guo et al. [68] show the feasibility for an adversary using a laptop outside a house to guess the browsing behavior of people living there. In the described scenario, the home network is protected by the Wi-Fi encryption mode. The attack is not based on breaking the encryption algorithm, but on extracting data from the signal which is composed by the sizes and timings of the packets. Since the elements of an HTML template are static over time such as CSS and image files in the form of headers or buttons, they are likely to show a distinctive pattern in the signal.

Some approaches to extract the knowledge of events and their internal links are described in Sun et al. [173]. Other works describe inference attacks that exploit background knowledge. For example, Guo et al. [68] present a localization and tracking algorithm for mobile target in an IoT environment according to information of neighboring sensor nodes.

2.1.2 Inference attacks from authorized acquisition

In recent years, new risks have emerged related to threats that come from trusted service providers/third party apps that are granted by the users' authorization to process their personal information [62, 110, 64, 131]. Alarmingly, studies show that third party apps usually ask for permissions more than they really require to run their service [35]. It was also stated in Zhou et al. [217] that apps can access the public resources of android (zero-permission) which could reveal private information. A detailed survey about privacy risks in mobile computing is provided in Haris et al. [72] including wearable devices.

We refer these new risks that come from permitted third parties as internal inference attacks. Ahmadinejad et al. [5] explained internal inference attacks based on the use of API and permissions-based access. Researchers analyzed the inadequacy of permission-based system to prevent inference of private data. For example, suppose a user does not want to share her birthday with Facebook application, but she is willing to grant the application the permission to access her "wall", since such permissions are needed by the application to deliver its functionalities. What she may not be aware is that this app may scan through her wall that looks for a day in a year in which have several postings of birthday wishes, thereby inferring her birth date information even though it was not permitted.

In addition, Carmagnola et al. [28] analyzed the possibility of identifying a user on different social networks and inferring private data by matching public data on the user profiles. Inferences were based only on data that the users defined as public on their privacy settings. The experimental study showed the possibility for an adversary to aggregate user data discovered on different sources into a more complete single profile and infer other data which are likely set as private by the user. It was observed that people typically ignore the idea that putting together data gathered from different sources makes it possible to obtain a very precise picture of their personal information, preferences and tendencies. Some risk factors were identified including some risky behaviors of users. The risk of information leakage found in OSNs is directly related with the number of shared personal data and the number of user profiles owned by the user over time. This risk can also exist on user data

coming from IoT applications. The higher the number of devices owned by the user that run in background applications (e.g., smart home gadgets, fitness appliances, etc.), the more historical data are collected which can lead to higher risks.

Recent practice of "lifelogging" (i.e., users' detailed documentation of their digital activities) and, in general, the use of personal IoT devices are increasing the privacy risks especially related to the possibility of using data mining techniques to infer personal and sensitive data from the shared data. Several examples are provided in Haris et al. [72]. Internal inference attacks have already been studied in database management (e.g., [43]), in online social networks and in approaches that use permission-based methods to access to user profiles [73, 5, 29]. Inferences can be derived from streams of raw data sensors and are able to discover mobility patterns (e.g. [209]), preferences and behaviors such as smoking and eating gestures [142, 46], health information [89].

Other internal inferences may be based on the identification of relations between users. Haddadi et al. [70] provided an example which shows the possibility for an allowed third party to infer whether two users have been spending time together, thus, being able to infer the location and mobility patterns of linked users who had not authorized the third party to access their location.

In this thesis, the main focus is on internal inference attacks where users have shared data to get a service (e.g., data acquired by sensors or inserted by the user and sent to third parties). These attacks are from authorized third parties that aim to infer further personal data that are not shared by users.

2.1.3 Measures against the inference attacks

This section cites some of the proposed measures against inference attacks. Techniques that are briefly discussed concern anonymization and transformation. The former aims to disjoint personal data from identifier data, the latter transforms the original data through generalization and perturbation to obfuscate the personal data.

K-anonymity [176] is a formal protection model against identity disclosure. The information released under the k-anonymity property guarantees protection of an individual's data. K-anonymity is ensured if personal data in the release are not

identifiable from at least k-1 individuals that are included in this release. This technique aims to protect against users' identity disclosure but does not protect against user's attribute disclosure [111, 29]. Moreover, in some cases, the identity of the subject is required.

Refinements of k-anonymity that address these issues are l-diversity [121] and t-closeness [111]. They use transformation techniques to reduce the granularity of data, thus, gaining higher privacy and protecting against the risk of attribute disclosure but reducing the utility of secured data. These techniques can be useful to manage sensitive data where privacy requirement is preferred over utility. However, they work on multiple dataset of different users and cannot estimate the risk of a single user data.

The technique described in Ahmadinejad et al. [5] introduces a protection model that transforms the original data into a form suitable to be shared with the third-party applications. This surrogate representation of the user profile is called a "view". The privacy policies stated by the user are used in these transformations. This view-based protection framework enables inference control by statistical correlation of data which blocks the attacker for inferring privacy disclosures. The level of transformation is ideal if the correlation between the sensitive and the transformed profile goes to zero. Each view consists of basic transformations and each transformation has to balance the goals of increasing privacy and preserving the utility of shared data. The main disadvantage of this technique is that heavy transformations of data often lead to loss of sensible information of the original data.

Some transformation techniques use semantic relationships among data. For example, Elkhodr et al. [48] describe a semantic obfuscation technique, called S-Obfuscation that secures users' location. The granularity of users' location information is adjusted according to the needed obfuscation. This technique consists of several levels of obfuscation and is based on a dynamic context analysis process which takes into account the user's privacy settings. Ontological classification of locations are used in the obfuscation. In the IoT scenario, wherein an information about a user's location van be very precise, this technique is highly relevant.

Finally, privacy inference graphs are also used to analyze privacy risks. For example, the algorithm described in Sun et al. [173] computes n-paths which are supposed to deduce privacy disclosures and the optimum strategies to block inference attacks according to such paths.

Measures against leaks and inference attacks usually consist in blocking or transforming the shared data. Transformations heavily depend on the type of data. They include generalization and perturbation to obfuscate the personal data being shared.

In this framework, the focus is not about creating a data transformation method but only providing the information of possible inferences. Therefore, the dependencies and correlation among data will be measured and will provide the risk information which will be discussed thoroughly in Chapter 5.

2.1.4 Inference Protection Methods

Inference protection methods consist mostly of static and dynamic analysis. Static analysis tries to identify privacy threats by automatically checking the code implementing an app, while dynamic and hybrid approaches analyze the behavior of applications as they are executed [215]. A further approach is by monitoring network communications and data flow to identify leakage of personal data. For example, *Haystack* [154] is an app that monitors the behavior of applications and performs traffic analysis under regular usage conditions. Traffic monitoring is also adopted in *ReCon* [155], a cross-platform application that detects Personally Identifiable Information (PII) leaks and allows users to block or modify PII shared over the network.

The ipShield framework [31] combines the approach of monitoring the accessed resources of the smartphone and recommending some actions by considering possible inferences of personal information. However, it's specific focus is on Android operating system and on the applications that use the phone sensors, while our aim is to protect against inferences in an IoT scenario where multiple apps interact with personal devices that are exchanging and sharing the user data, i.e. data from sensors and from the phone resources (accessed on user permission), and data explicitly provided by the user on request of the app.

Recent approaches face the issue of balancing privacy and utility of the service being provided by empowering users with control over the release and the processing of personal data. For example, Ardagna et al. [7] proposed a model to continuously release user location information with the option of obfuscating some of them under specific conditions, which is based on the user's behavior and explicit preferences.

Differently from the dynamic analysis and traffic monitoring, our approach tries to identify privacy risks before the application is executed. Moreover, the main difference with these approaches is that our focus is on inference prevention (apriori) while most of the works mentioned above aims to detect and reveal PII leaks. The framework informs users which unshared data can be inferred by an allowed third party, given the set of their shared personal data. This happens during installation which means that prior information is already given to the user before the damage.

2.2 Privacy Preference Modeling and Recommendation

In this section, related work regarding the privacy preference modeling and recommendation are mainly discussed. First, privacy management and frameworks necessary for IoT are introduced, together with a short discussion on the birth of privacy preference modeling and recommendation. This is succeeded by a discussion regarding privacy preference modeling using ontologies and machine learning. Then, a brief overview of user privacy behaviour is discussed briefly. Finally, the laws and principles regulating users' privacy are briefly discussed, which are taken into account in the framework.

2.2.1 Privacy Management

Mobile privacy permission systems have been well-studied in literature. However, they do not cover the scope of the new IoT devices, such as fitness trackers, that expand and extend the services and personal data that must be managed. In this thesis, the capabilities of mobile privacy managers will be extended to cover the requirements in the Internet of *Wearable* Things (IoWT) [138] context. Background

information about privacy permission management in mobile systems will be introduced since they served as groundwork for the IoT context. Then, the related work regarding privacy management in the IoT will be discussed.

Permission Management in Mobile Systems

Previous studies in mobile privacy (e.g.,[55]) have proven that mobile interfaces lack the potential to provide the necessary user privacy information and control for both Android and iOS systems [115]. Several solutions from literature have been proposed from then on to improve mobile privacy protection and offer users more privacy control (e.g., [17]). These leads into rapid improvement of privacy management of current mobile systems (i.e., from Android 6.0+ and iOS 5.0+), providing more control on the user's privacy settings.

Android permission systems can be mainly categorized as Ask On Install (AOI) and Ask On First Use (AOFU) privacy models [182, 195]. 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 clearly gives less privacy control. Also, only few users read and pay attention to the install time permissions, and even fewer than this understand their meaning [55, 87]. These issues made room for TP apps that manage app privacy such as Turtleguard [182] and Mockdroid [17].

On the other hand, the AOFU model [182] (Android 6.0 and above) only asks permissions during the first use of an app and when an app uses a specific feature that needs the respective permission. In this case, the user grants the permission during the actual provision of the service and will be able to weigh his willingness to share against the utility of the app. The user 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 as the previous model does not allow users to be informed effectively [59]. Moreover, it has been proven that interactive notification is more efficient in informing users request access [59]. It is noteworthy to discuss these two models as currently, 34% of the Android users are still using the AOI model[2].

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

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

In terms of privacy management, iOS system has used the AOFU model for location permission since iOS version 5.0[3], with a more comprehensive roll-out in iOS 6.0 and onwards [6]. Although it is not open-source like Android, it does not stop researchers from finding ways to improve the privacy settings. For example, in [4], *ProtectMyPrivacy* is an app specifically designed for jailbroken iOS devices that preserves user privacy by substituting anonymized data for user's private data. Although jailbreaking is deemed legal, it is not advisable to do so as jailbroken devices can be used to install pirated apps that can risk user privacy [4]. Privacy managers for unjailbroken iOS devices also exist, however, it obviously has reduced functionality. For example, *PiOS* [47] is a Privacy Manager which only has the function to check if the installed iOS apps have committed privacy breaches.

According to Tsai et al. [182], AOFU model can be extended to Ask On Every Use (AOEU). This ensures that user privacy preference can be accurately enforced and dynamically set during the actual request of access. However, it is not usable in practice since applications request access with great frequency, averaging at once every quarter of a minute. This leads into an impractical solution that would be ineffective in practice. This is also confirmed by prior works that have runtime prompts for every sensitive request [194].

Another privacy related issue in mobile systems is that users are not fully aware that apps can access sensitive data even when not in use [178]. For this reason, Tsai et al. [182] developed TurtleGuard, which decides whether to give or deny access of apps whether they are actively being used or not. This is one of the most important contextual factors which have been addressed in literature [194].

Frameworks for Privacy Management

Initially, privacy management frameworks started from mobile applications. For instance, the aforementioned ipShield [31] is a context-aware privacy framework for mobile systems that provides users with great control of their data and inference risks. Frameworks for general privacy management also exists in literature. For instance, My Data Store [187] offers a set of tools to manage, control and exploit

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

personal data by enhancing an individual's awareness on the value of their data. Similarly, Databox [33] 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. Our proposed framework include the idea of developing a personal data manager for privacy management in the advent of IoT.

2.2.2 Birth of Privacy Preference Modeling

In 2001, Kobsa et al. [99] suggest that privacy settings should be dynamically tailored to both legislative rules and the individual user needs since different factors affect user preferences. Given the diversity of user preferences, context conditions and regulations, privacy preference modeling becomes a challenging task.

In ubiquitous computing environment, privacy modeling is tackled since early 1990's [14]. In the early 2000's, Kay et al. [85, 86, 24] focus their research on supporting user scrutiny and control over the information held by applications. For instance, *Personis* [85, 86] 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, Brar et al. [24] introduced Secure Persona Exchange (SPE), which 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[4] vocabulary to provide tools for representing and storing user preferences as subsets of user model(i.e., *personas*) each intended for use by particular applications. Though the P3P became obsolete due to its lack of adoption, most of its main concepts are still used for data protection regulations.

Context-aware privacy modeling have proven to enhance the accuracy of user's privacy preference prediction [195, 108]. Context is defined as the circumstance (e.g., what, when, who, where, how, etc.) under which a TP application requests access to data. The context allows to enhance the prediction, for example, "when" and under "what" circumstance plays a big role in predicting user preferences [195].

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

Using interviews and online surveys to model the privacy preferences of potential IoT users, Lee et al. [108] identified the contextual parameters that have the strong influence on the user's privacy preferences. These parameters include the type of monitoring, the type of information collected, the entity collecting the information, the frequency of monitoring, the location, and the reason for the collected data. The identity of the information requester (*who* context), which is an important determiner of user's privacy decisions as stated in previous studies, was also confirmed in this study.

Leveraging the dataset collected by Lee et al. [108], Bahirat et al. [11] created a privacy-setting interface that allows users to deny/allow IoT devices access to their personal information. They also modeled users' decisions as a means to come up with default privacy profiles.

Preference modeling was also explored for privacy and social recommendation of social networks. For instance, Facebook users are found to have 6 types of privacy profiles which range from Privacy Maximizers to Minimalists [95, 197]. In Wu et al. [203], 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) enhance the modeling of user preferences.

In the health/fitness domain, emerging sensors and mobile applications allow people to easily capture fine-grained personal data related to long term fitness goals. Focusing on tracker data (i.e., weight, activity, and sitting), Brar et al. [24] discover that user's preferences change for every sensor (i.e., weight being the most important [24]). Also, their study concludes that users want to have control and have a personal copy of their fitness data.

Modeling the privacy preference about location received great attention in the literature given the sensitivity of this information [8, 188, 204, 6]. For instance, Assad et al. [8] study the user preferences about the release of location information and provide support to differentiate the release. In Vicente et al. [188], not only location but also absence and co-location privacy are considered. In Xie et al. [204], location sharing privacy preferences are studied with respect to different groups and different contexts, including check-in time, companion and emotion. These studies confirm that users want to control the privacy of information and this is specifically

important in ubiquitous environments.

2.2.3 Privacy Preference Recommendation

Enhanced permission settings surely give more control to the user, however, as the number of applications that the users utilize increases (averaging 35 apps/user [66]), the number of permissions in a single application increases (averaging 5 permissions per app[5]), and the number of devices that the users own increases (averaging 4 devices per user[6]), these permission models will not be enough. The burden of individually setting each permission will become a tedious task for the users, which also make them prone to errors [122, 108, 2]. Furthermore, users are increasingly unable to make decisions about privacy settings due to limits in their available time, motivation, and their cognitive decision-making abilities [108, 2]. In this section, we describe some of the approaches that have been proposed to address this problem.

Privacy nudging is an effective method to increase user awareness [6]. Nudging allows users to be informed and aware on both their privacy settings and how TP applications access their data [117, 59]. In the study conducted by Liu et al. [117], their results report that 78.7% of the privacy nudges were adopted by the user. However, it does not mean that if they adopted the recommendation, they have fully understood or it really has something to do with their privacy perception. Nudges often inform users of the possible breaches even though it may not concern their privacy preference. Privacy nudges lack personalization and only provide generalized recommendation.

Another approach that is more user-centric is the user-tailored privacy [95]. It models users' privacy concerns and provides them 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 the right amount of both the privacy-related information associated to them and the useful privacy control that do not overwhelm or mislead them. However, in practice it is hard to implement

[5] Average number of permissions per app:http://www.pewinternet.org/2015/11/10/apps-permissions-in-the-google-play-store/

[6] Average number of devices per consumer: https://blog.globalwebindex.com/chart-of-the-day/digital-consumers-own-3-64-connected-devices/

general privacy model as the idea is too broad and abstract especially in the diversity of privacy perception of users.

To provide recommendation, the approach in Lin et al. [115] and Liu et al. [117] is to initially create a set of predefined privacy preference settings and select which can be recommended to the users. This can be attained by using Machine Learning algorithms to predict the best-suited preference settings for the user. It shows that in a diverse sea of permission settings, there exists some profiles that could collectively describe such diversity. These privacy profiles are collections of settings that correspond to privacy preferences of similar-minded users [115, 118, 117, 94, 93, 197, 204]. For instance, Liu et al. [118] identified six privacy profiles based on the analysis of 4.8 million users' privacy settings. Liu et al. [117] extended their work by adding new features, such as the purpose of information and app categories, in modeling user privacy profiles with the inclusion of privacy nudges that makes users aware of unexpected data practices from third parties. This approach can provide decision support for privacy recommendation given that a privacy profile that best describes a user can be found [117]. This recommendation approach is adopted in this thesis and extended in the IoT context.

2.3 Privacy Preference Modeling using Ontologies

One of the main problems of IoT is its heterogeneity. IoT devices must be interoperable for the IoT paradigm to work. Yaqoob et al. [210] state that IoT has three main interoperability challenges, namely technical, semantic, and pragmatic. Technical and pragmatic interoperability refer to the device capabilities (i.e., standards and protocols) and TP intentions, respectively. Semantic interoperability is a requirement to the machine computable logic, knowledge discovery, and data federation between information systems. For our PDM to work in the IoT context, we take into account semantics in the PDM.

Semantics is one of the standardization opportunities that can be implemented which makes easier for the service providers and consumers to work with an IoT

ecosystem [165]. An ongoing standard specification, oneM2M (M2M: machine-to-machine)[7] is developing a semantics enabler to bridge the gap between current IoT resources [165]. OneM2M states that their platform is enhanced with semantic capabilities to ensure cross-domain interoperability between IoT devices and applications at the level of communication and data [139].

In literature, there exist a number of privacy preference modeling frameworks that use the semantic web approach. PPO [157] has been the pioneer for modeling user's privacy preferences, giving users' fine-grained control of their preference.

P3P (W3C Platform for Privacy Preferences) can be considered a reference model for automatic processing of privacy preferences. Users can express their preferences, and their browser warns them if a site does not meet these preferences [13]. Many proposals about privacy enhancing technologies are based on P3P, which did not became successful due to some usability issues and the lack of enforcement when it was defined. Additionally, the advent of social networks and IoT brought new complex privacy requirements.

In this work, we propose an ontology, which is used by the PDM, specifically for privacy management in IoT context. Privacy preferences in the IoT (PPIoT) context critically depend on the reason for data collection, the persistence of access, the location, the retention period and the method of usage [108, 78, 141, 21]. These aspects constitute the requirements for privacy management in the Iot paradigm [120, 184] and are taken into account in the proposed PPIoT Ontology. They also ascertain that PPIoT is compliant with the GDPR, which enforce the requesting entity to clearly specify the reason, usage of data, frequency and method of data collection, and the retention period of the collected data.

Below, we first describe the base ontologies that are extended by our PPIoT ontology and then we report about other related ontologies for privacy modeling. The full details of the PPIoT ontology will be described in Chapter 6.1.

2.3.1 Base Ontologies

In line with best practices for ontology reuse, our PPIoT ontology integrates the current Privacy Preference Ontology (PPO) and the W3C Semantic Sensor Network

[7] http://www.onem2m.org

(SOSA/SSN) Ontology. The PPO was created to aid users in managing their privacy settings in the realm of linked data [157]. Later, it was extended to facilitate social network applications. The PPO allows users to have more fine-grained control over their personal data. One of the main features of this ontology is the possibility to set multiple privacy preferences each with a different set of conditions for a single user.

The original W3C Semantic Sensor Network (SSN) Ontology was aligned to the DOLCE-UltraLite3 Ontology and was based on the core concepts of the Stimulus–Sensor–Observation ontology pattern [39]. Due to the rapid expansion and diversity of data and its providers, it has been improved and is now based on the Sensor, Observation, Sample, and Actuator (SOSA) ontology pattern [71] to include broadened definitions (e.g., social sensing applications). It has low interoperability fall-back and focuses on enlarging the target audience and its application areas. As of October 2017, it became a W3C Recommendation.

2.3.2 Ontologies for Privacy Modeling

A survey on ontology-based privacy modeling can be found in Perera et al. [146]. Below, we briefly describe the ontologies that are most related to our study.

The ontology described by Zhang et al. [214] shares similarities with PPO and is focused on defining privacy rules. Each rule must contain a data class and a conditions class. The main features of the conditions class include the duration, purposes, and recipients of collection, the period on keeping the collected data, the user's privileges, and ways of handling disputes. This ontology was intended for applications of context-aware systems.

PROACT is an ontology that models privacy in relation to tasks and user activities [141]. The authors introduce the concept of "activity sphere", which is a temporary abstract space defined to limit the incoming and outgoing information. PROACT is used to define privacy policies based on restrictions and rules for accessing and using each resource within an activity sphere and the information the resource collects and manages.

The authors of the PPO also created a light-weight ontology, named Privacy Preference Manager, which is a semantic representation of a tool that allows users to

deny/allow access to their data based on the Web Access Control (WAC) vocabulary [157].

The privacy preference model proposed Bodorik et al. [21] is characterized by regulations and conditions that are specified by the user. Conditions can concern the purpose of the data recipient, usage and retention, disputes, remedy, and access control. The model also specifies the properties of a steady set of user preferences and their maintenance operations.

Hu et al. [78] propose an ontology using privacy rules. Their ontology, which also follows the FIP guidelines, is intended to capture allowed/denied purposes of data collection, allowed/denied access for individual entities, retention period, obligations, policy, and action. This enables the creation of global rules to define preferences for higher-level conditions, such as giving a recipient access to data that came from medical applications, even if some of the needed parameters are not defined.

Setting rule priorities is another feature of privacy preference management. Rei is a policy language that aids users in expressing their privacy preference conditions with a priority hierarchy [83]. This level of expressiveness is a step forward in the enrichment of privacy specifications, as it helps in resolving conflicts. This is especially relevant in the context of IoT, where several conflicting conditions can occur. However, Rei lacks the power of negotiation, since it can only set multiple conditions on the user side (i.e., it does not consider the TP side).

The usefulness of modeling trust is thoroughly described in Iqbal et al. [79]. Furthermore, Martimiano et al. [124] model the trustworthiness of TPs using principles similar to Friend Of A Friend (FOAF). They extend this principle by defining fixed sets of classes with predefined assignments on the level of trust (e.g., close family, friend, work mates, unknown).

The ontologies mentioned in this subsection contain unique features that allow users to be more expressive regarding their privacy settings. However, these ontologies are limited in their application domain and focus exclusively on the user side, not on the TPs. To extend the ontological approach to privacy management to the field of IoT, it must have room for negotiation between the user and the TPs [146]. Our proposed PDM takes this into account.

2.4 Privacy Preference Modeling Using Machine Learning

Using Machine Learning (ML) for modeling permissions was deemed a practical and an effective approach [182, 195, 117, 118, 108]. Thus, prediction of user privacy preferences is well-studied [42, 158, 53, 18, 20, 216, 19, 107].

By training contextual user data and a subset of users' prior privacy behavior, ML approaches in Tsai et al. [182] and Wijesekera et al. [195] achieve significantly smaller error rates over the current AOFU model and reduce (but do not eliminate) user involvement. AOFU model is criticized for asking only once during the initial utilization of the app when the user's preference is different under which it subsequently requests access. Users can still revisit this setting in a user interface and check the decisions made by ML.

Frank et al. [57] conducted one of the earliest study on mining mobile app permissions. The study aims to mine patterns in Android app permission request by using matrix factorization techniques. They found over 30 patterns of TP app requests. Furthermore, Lin et al. [115] identified patterns in user privacy preferences for a given set of preferences asked in mobile apps for different purposes of data access.

Liu et al. [118] showed that a few privacy profiles can be derived from a sea of diverse settings. A total of 239,000 privacy settings were studied for 12,000 different apps. In addition, their system prompts user if it is uncertain (i.e., low level of confidence) during the prediction of privacy profiles for a given user.

For social networking apps, Sadeh et al. [158] used ML such as Random Forest to provide automated privacy decision on behalf of users. Also, Fang et al. [53] provides a system for social networking services that infers access control policies through supervised learning by iteratively asking users questions about their sharing preferences with their friends. Using 45 Facebook users, the burden of setting privacy was effectively reduced by the system. It also reached 90% accuracy in predicting personal privacy policies using only a subset of labeled training data. A privacy prediction model was introduced by Dong et al. [45] for social networks, which is established from psychological principles. In addition to predicting user preference, the model also provides personalized advice to users regarding their privacy

decision-making practices.

A privacy-preserving information sharing platform named SPISM [19] was deployed on Android OS. SPISM assists users in making decisions for disclosing personal data on different levels of granularity. SPISM claims to have 90% accuracy on predicting users' privacy decision by using a logistic classifier. Furthermore, a Personalized Privacy Assistant (PPA) [117] was also used for android applications to support user decision. Privacy profiles were developed using hierarchical clustering which are groups of similar privacy user attitudes. They extended in to prediction of user decision during permission request of third party entities by using SVM classifiers. Furthermore, PPA could also provide nudging to support user decision.

Probing users' privacy preferences by asking them a small set of general questions is typical in recommender systems. It helps in recommending the default settings or default profile for a certain user. For instance, a few questions are asked in the setup phase of the Locaccino system [196] to guide users through the default location sharing profiles suitable for them. Similar approach are used in [98, 153, 128] for setting up users location sharing rules.

Privacy profiles are only estimates of the users preference, given a diverse sea of permission settings. Further personalization would make it more accurate. The chosen or recommended privacy profiles can be the initial set-up and would be refined by the user or further machine learning approaches [41, 128, 88].

Our proposed PDM also uses ML techniques to model and recommend the privacy preferences of the users. This PDM recommendation aims to reduce the need to manually set the user's permission settings.

2.5 User Behavior on Privacy Preference

Literature suggests that contextual cues can be used to detect privacy violations [134, 12]. Thus, context-based privacy preference are studied [195, 108]. However, users usually are unable to provide such privacy decisions given their lack of time, motivation and their cognitive decision-making abilities [2, 169].

Benisch et al. [15] also show that users felt more secure in privacy settings that are comprised of contextual factors rather than the traditional whitelisting. Some of

the contextual factors that are discovered by the authors that have significant impact on location privacy includes the actual location, time, and the day of the week.

Another issue that needs attention is the user's certainty on his preference. Users themselves have been reported that their real preference is inconsistent with their actual information disclosure behaviors. This phenomenon known as the "privacy paradox", which is well-established in previous research [135, 44, 186, 170, 207]. In understanding privacy paradox, Dinev et al. [44] reported that perceived privacy risk and privacy concerns are two factors related to the willingness to provide personal information in Internet e-commerce transaction. Therefore, it should not be assumed that disclosure behaviour reflects a lack of concern with respect to privacy. In Internet e-commerce, the strong relationship between perceived Internet trust and willingness to provide personal information suggests that trust is an important condition for completing online transactions. Therefore, trust is a factor for the actual information disclosure regardless of their real preference. However, the study concludes that more research needs to be done to understand completely the privacy paradox.

Users also behave differently if they know they are being monitored. In the study of home IoT in Choe et al. [36], users are less likely to share some activities (e.g., intimacy behavior, cooking or eating) at home when various sensors are installed. Some areas inside the home are considered more private than others, which is an important factor for the deployment of home IoT devices.

Furthermore, users usually do not pay enough attention to privacy permissions. Most users do not really read (only 17%) nor understand (only 3%) the permissions and are very susceptible to install apps with look-alike name to other apps [55]. This highlights the need to make privacy permission clearer and well-modulated and give users informative tools to control their data and substantial information on how their data are being processed [30, 40].

Attribution mechanisms have been proven to be a necessity to help users better understand smartphone application resource access [178]. When a system change occurs, it may be helpful to place attribution information in places where users can undo those changes. Thompson et al. [178] have found that this information is expected by users to be in the device's Settings app. The study also shows that only

very few of the of participants (i.e., 22%) know that apps can still run in the background and they have the same privilege as if being actively used. For this reason, user expectation has been studied by Wijesekera et al. [194] through regulating access based on being used or not since users perceived this permission request from background as unexpected or inappropriate.

The idea of supporting users in decision-making on their privacy is becoming a relevant research field, given the increasing complexity of privacy settings and data sharing models [3, 192]. This reflects also the principle of empowering users with more control on their data, which has given rise to several proposals and tools for managing personal data. Examples include the personal data manager described by Zyskind et al. [218], My Data Store [187], and Databox [33].

Privacy management is closely related to the studies on privacy perception and user behaviors. Several researches show that users have different perception of privacy value and have different behaviors in managing their privacy preferences [197, 72]. The disclosed behaviors have been described as multidimensional by Knijnenburg et al. [93] since users seem to differ for the kind and the degree of information that are being disclosed. This suggested the idea of adopting personalized techniques to tailor the management of privacy settings [198].

An analysis of the factors that influence privacy behaviors specifically with respect to mobile applications is provided in Haris et al. [72]. The authors show that besides personal differences in perceiving the value of privacy, behaviors are often influenced by the fact that many users are not aware of their released data. This may be due to the lack of understanding of requests about mobile permissions (often vague, confusing, and poorly grouped), lack of knowledge about the number of sensors that can release personal data, and unawareness of possible integration of harmless data to infer private data.

Other factors reported that can influence the user's behavior toward privacy are the type of information, the retention period, and social aspects such as popularity and recommendation by friends. Based on these factors, crowdsourcing could be a useful approach to understand the users' perception of privacy [114] to simulate different scenarios that includes these factors.

For this reason, our proposed PDM aims to support the users' decision and make

them aware of their released information. Together with AID-S, the decision support also includes the computation of risks on the probability that a third party can potentially infer further user information. To further capture users' diverse perceptions, the crowdsourcing approach was utilized as pointed out in Lin et al. [114] in understanding users.

2.6 Lawful Privacy Protection

User privacy is also protected and regulated by laws and principles toward data handling. The main privacy principles and laws governing user privacy are briefly discussed in this study.

2.6.1 Digital Rights Management

Digital Rights Management (DRM) mainly concerns on the publishing and control of the consumption of commercial and digital media and prevents against illegal re-distribution. It also fights against re-distribution and illegal acquisition of user data [104].

Kumar et al. [104] state that privacy policies can be defined for personal data rights object or license instead of implementing commercial media principles which are negotiated during data processing. User's awareness and permission can be used instead to prevent illegal distribution of their personal data.

2.6.2 Fair Information Practices Principles

Fair Information Practices (FIP) principles are long-standing guidelines regarding the collection and use of users' information that aim to protect their privacy [63]. The FIP principles are transparency, individual participation, purpose specification, data minimization, use limitation, data quality and integrity, security, and accountability and auditing. The inclusion of the FIP principles in privacy frameworks is much needed, especially in the management of IoT user data collection [120, 184, 106] as recent reports of privacy breaches of supposedly trusted TPs are increasing [106].

2.6.3 General Data Privacy Protection

As of May 25, 2018, the European Union (EU) enforce the General Data Protection Regulation (GDPR) [61] that applies to the storage, processing and use of the subject's personal data from the accessing third parties which may or may not have been established in the EU as long as they operate in an EU market or access data of EU residents. It requires users to provide explicit consent to privacy options expressed by third parties. This results in a complex task for the users given the number of devices and applications which have to be read and processed specifically.

After the GDPR implementation, an ontology was proposed in [49] to represent GDPR rules concerning Cloud data. However, the ontology does not take into account user privacy preferences and concentrates only on the obligations of both consumer and provider of cloud data.

The principles that relate to personal data protection are explicitly stated in Article 5 of the GDPR Regulation: Lawfulness, fairness and transparency; purpose limitation; data minimization; accuracy; storage limitation; integrity and confidentiality. Broadly speaking, the GDPR requirements concern two main issues:

- the management of personal data from the TP (data processing, sharing and storage);
- The communication between the TP and the user about the management of personal data (transparency, controllability, accuracy).

In this thesis, the proposed privacy framework complies with the GDPR through the PPIoT ontology, which is designed to include classes and properties that address the GDPR requirements for the management of personal data. It includes the reason, method, purpose and persistence of data access, and the maximum retention period of data in the hand of the accessing parties.

Chapter 3

PerNANDO Framework

3.1 Research Questions and Methodology

Based on the discussions of Chapter 2, managing users' privacy and protecting them from inference risks are highly needed in the IoT context, which also need to be aligned with the mandated GDPR. This thesis aims to provide a framework for privacy that focuses on these concerns.

To this point, our framework aims to solve the following specific research questions:

> **RQ1.** *How can we protect IoT users from potential inference of their private data given their disclosed data?*

To answer this question, we aim to realize an inference discovery system that can compute risk of data correlation. We utilize a probabilistic approach of measuring privacy risks through Bayesian Networks. The risk probabilities are then informed to the user to support their decision in granting access to the third parties. This will be answered in Chapter 5.

We created our inference system using the real user data on fitness IoT (i.e., Fitbit Trackers) from the Open Humans Foundation[1] and the crowd-sourced Fitbit dataset generated from the respondents of a distributed survey via Amazon Mechanical Turk [2] by Furberg et al. Our quantitative analysis and the evaluation of our risk computation are explained in Chapter 5. Currently, our results are valid only for fitness IoT and/or domains that use similar/subset of fitness data from fitness trackers.

[1] https://www.openhumans.org/

[2] https://zenodo.org/record/53894#.W-76oehKjIU

RQ2. *How can we model user privacy preferences in the advent of heterogeneous personal IoT data?*

To answer this question, the proposed framework utilizes the semantic web approach to solve the problem of heterogeneity. As of the time of writing, semantic modeling of user privacy preferences in the IoT paradigm is not yet defined. This work proposes PPIoT Ontology, which is based on PPO and SOSA/SSN Network that extends semantic modeling of privacy preferences. This will be answered in Chapter 6.

The proposed PPIoT Ontology is currently limited to the Fitness IoT domain. The terminologies are based on the main fitness trackers that are used as of the time of writing. PPIoT Ontology also complies and uses GDPR terminologies. Chapter 6 will explain all the details regarding the proposed Ontology.

RQ3. *How can we aid users to set their privacy and provide them with suitable recommendation?*

To answer this question, we created different privacy profiles using ML clustering. Then, we use supervised ML to find determiners that can tell which privacy profile best fits for a given user. Finally we provide different recommendation strategies that interact with the user to provide the recommendation. This will be answered in Chapter 7.

To give more details for RQ3, the overview of the methodology is depicted in Figure 3.1. User Profiling & Recommendation can be divided into 3 main approaches. First, modeling privacy settings that can generally represent privacy preferences among IoT devices has been defined. This allows to answer RQ2, as shortly explained above.

Then, using Machine Learning (ML) techniques, we select predefined privacy settings, called privacy profiles, which are able to represent most of the IoT users' privacy preferences. These profiles range from privacy-aware to unconcerned users which are developed using ML clustering. Then, we find determiners that can classify a user to which privacy profile he/she belongs. These determiners include some questions regarding some privacy items, situational and social questions, which are developed using ML classification.

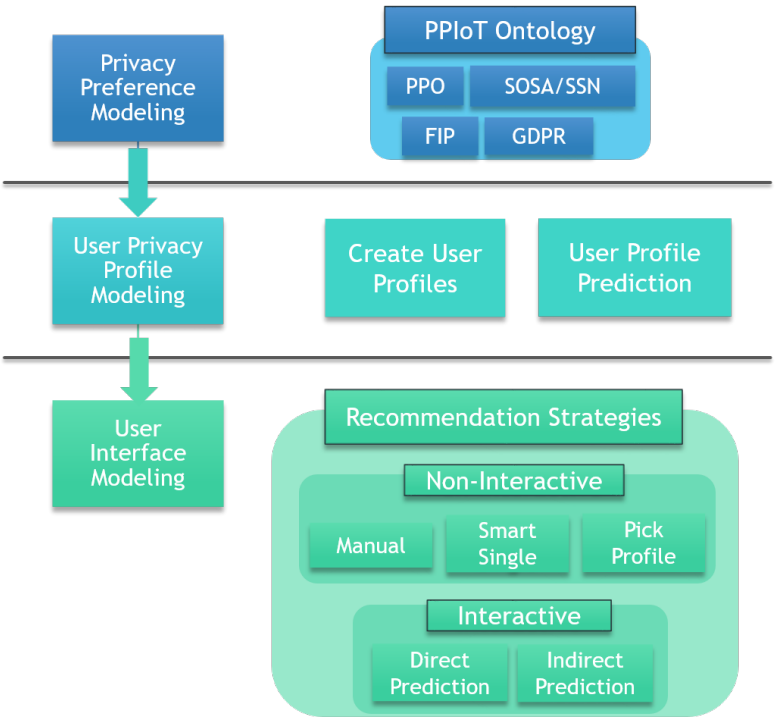


FIGURE 3.1: The workflow for User Profiling & Recommendation.

Finally, we develop recommendation strategies that interact with the user. These strategies can be mainly divided into interactive and non-interactive User Interfaces (UIs). Interactive UIs ask users questions that enable the system to classify which profile a user belongs based from the results of the ML. On the other hand, non-interactive UIs do not ask any question at all. These strategies answer our RQ3.

For RQ3, the methodology includes creating a mock-app simulator that simulates an environment that set-up the setting of privacy preferences of a fitness app. We run the simulator on real fitness tracker users (i.e., Fitbit users) from the Amazon Mechanical Turk, which enables us to collect dataset for real privacy settings. This also means our result are limited to such domain. The analysis and evaluation of these methods are fully explained in Chapter 7.

For this study, our framework is applied and tested on the fitness domain, however, the methodology could be extended to other IoT domains as well. The fitness domain is chosen in this study since, as of the time of writing, these IoT devices collect most of the users' personal data, which is well-documented in Chapter 4.

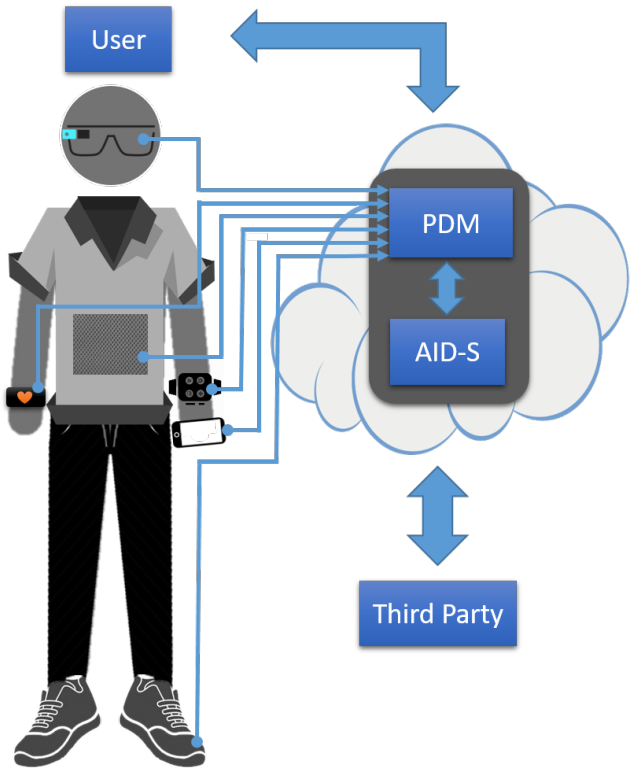


FIGURE 3.2: The current personal IoT scenario managed by the PerNANDO framework.

3.2 Framework Definition

In this chapter, the proposed Personal, Non-intrusive, and App-Neutral Data Organizer (PerNANDO) framework will be defined. This framework aims to be non-intrusive, as the users' data are not kept and stored. Only the description or user's preference of sharing of such data are needed. It is App-neutral since it aims to be compatible on both regular and enhanced types of third parties, which will be discussed in-depth in Chapter 6.

The overview of the framework is shown in Figure 3.2. It depicts a modern person having heterogeneous personal devices and things capable of connecting to the network (e.g., smart glass, smart watch, fitness tracker, smart fabric, smart shoes and smart phone). In this scenario, the user's IoT devices are connected to his PDM inside the *PDM+AID-S* function block.

The interaction between the user and the IoT device (handled by its manufacturing third party application) is managed by this function block. The framework acts as a gateway of the information flow. The request of personal data from the third

TABLE 3.1: The PerNaNDO Work Package Composition.

Functionalities	Tasks	Packages
Dialog Management	Interaction among entities	PDM
Access Control	Policy Statement evaluation	
	Authentication & Authorization	
User Profiling & Recommendation	Privacy preference modeling	
	Privacy preference clustering	
	Privacy preference recommendation	
Inference Discovery	Inference risk computation	AID-S
Risk Recommendation	Optimal privacy setting	
	Transformation	

party, which is in a form of a **Policy Statement**, is regulated by the user's PDM before giving the authorization. It checks if the Policy Statement is coherent with the user's privacy settings.

The AID-S, on the other hand, computes the inference risks associated to data disclosure. This combination enables to ensure full protection of the user concerning privacy risks. The framework also ensures that it respects the GDPR principles regarding data handling and informed consent of the user. The components of the framework and the workload distribution are stated in Table 3.1. The table shows the core functionalities provided by the framework and the related tasks within the PDM and AID-S packages.

3.2.1 PDM Tasks

This section defines the tasks that are managed by the PDM as displayed in Table 3.1. They are based on current PDMs in the literature (e.g., [33, 218, 187, 177]) and are extended to work in the IoT context.

Dialog management

The PDM is the gateway between the user, the third party and AID-S. The dialog management task is aimed to handle and coordinate the communication among all these entities.

- PDM2User. The functions of this subtask are to receive the user requests for new loT device installation, manage the interactions with the modules that are in charge of profiling the user, and inform the user with risk recommendation from AID-S.
- PDM2ThirdParty. The functions of this subtask are to receive the third party application request (consisting of the Policy Statement), and coordinate the interaction with the Policy Statement evaluation task and the negotiation with the third party in case the Policy Statement does not satisfy the user privacy requirements.
- PDM2AID-S. The function of this subtask is to handle the requests from PDM to AID-S for inference discovery and handle the requests from AID-S to PDM for privacy preference setting. Moreover, it receives the results from AID-S which consist of recommendations to be provided to the third party and to the user in asynchronous phases.

The most crucial communication rises from PDM2ThirdParty. Thus, this work focuses on the realization of PDM2ThirdParty by using the Semantic Web approach and how it can perform negotiation through the proposed PPIoT Ontology, which will be discussed in Chapter 6.

Policy Statement Evaluation

PDM evaluates a request of a third party in a form of a Policy Statement. In current PDM models, Statements consist of information that are essential for describing which data will be collected, stored and processed by the third party, their quantity, frequency of acquisition, retention period of the accessed data, and purposes. These principles are from the GDPR handling regulation on personal data. The main goal

of this task is to evaluate if the third party request respects the privacy setting of the user which will be shown in Chapter 6.

Authentication and Authorization

A basic function of PDM is to manage access control. This includes both the user and the third party. Concerning the user, this task is needed when the PDM is deployed as a service running on a remote server or in hybrid client-server models that request the user's consent.

Concerning third parties, Authentication is aimed to manage a list of authenticated third parties and use authentication protocols to secure the transactions with them. It is also possible to refer to a Certification Authority (CA) regarding the status of the third party.

Mainly, the framework does not have the capability for authentication as it is a security-related feature. This work task can be completed using literature solutions, which heavily focuses on this security-related privacy, and will not be taken account in this study as discussed in Chapter 1.

On the other hand, Authorization of third parties is granted after being certified and if their request has been successfully accepted. PDM Authorization can grant access to TPs if the user's privacy preference conditions are met. Chapter 6 will show a use case of how PDM can authorize accessing TPs using the PPIoT Ontology.

User Privacy Profiles and Recommendation

To evaluate third party requests, an important task managed by PDM is creating a user profile with the user's preferences about privacy. The user profile is based on the user's privacy preferences that are recorded and stored. This enables the PDM to provide comparison with the TP Policy Statement and check if it can be granted access.

Privacy preference settings represent the user's perception about confidentiality of his own personal data items. For different levels of privacy, i.e., not only allow or deny a privacy item, an example of user threshold setting can be found in [31] in which a specific user has an assigned priority vector,

$$priority = (p_{l_1}, p_{l_2}, \cdots, p_{l_n}) \tag{3.1}$$

where priority means confidentiality and each element, $p_l \in \{0, \cdots, p_{max}\}$ expresses the user's priority level set for each category, l, in numerical value having p_{max} as the maximum (i.e., maximum confidentiality).

The PPIoT Ontology have the capability to let users express a more fine-grained privacy preference, which will be shown in Chapter 6.

For this study, the privacy profiling using ML will be elaborated thoroughly together with strategies for recommendation which will be discussed in Chapter 7. Different privacy settings are created to alleviate the burden of manually setting several permissions. Then, the profile that best suits a given user will be recommended by PDM.

3.2.2 AID-S Tasks

AID-S is portrayed as an external function that the PDM can access. Conceptually, they are separated since PDM work packages focuses on the management of user preferences.

The main functions of AID-S are to predict an inference risk based on a set of input data, and to provide solutions to reduce the inference risk. The Inference risk computation task starts on the PDM's request regarding possible inference attacks of a third party in the case its Policy Statement is accepted.

Inference Risk Estimation

The objective of this task is to compute the risks that a third party may infer user's private data given the available data released to it. An ideal representation of inference probabilities based on dependencies among user data, independently of shared and non-shared data, is the Inference Matrix, I, where:

$$I_{i,j} = \begin{pmatrix} P(a_1|c_1) & \cdots & P(a_1|c_j) \\ \vdots & \ddots & \vdots \\ P(a_i|c_1) & \cdots & P(a_i|c_j) \end{pmatrix} \tag{3.2}$$

- $I_{i,j}$ is the matrix containing all the probabilities of inferring an attribute, a_i , given the dependencies of a set of other attributes c_j;
- $a_i \in D(1 \leq i \leq |D|)$, where D is the user Data Set
- $c_j \in P\{D\}(1 \leq j \leq |P\{D\}|)$ where P{D} is the Power Set of D
- $P(a_i|c_j) \in [0, 1]$

The computed probabilities depend on the probabilistic model or learning algorithm that can be deployed (e.g., RST, KN, Bayes Filter, HMM, etc.). It should be noted that the method is open to any probabilistic model that can be found in the literature.

For example, the study in Cai et al. [27] use RST, Naive Bayes and KNN to study the inference in social networks (e.g., Facebook). In these terms, the matrix can be ideally used to represent all the possible inferences of personal data based on user data correlation.

Based on this matrix, the inference risk for a given user can be computed by considering the user's specific privacy preferences that are used as thresholds, t_i for a_i. The Inference Matrix is an abstract representation but several algorithms are available to compute inference measures for limited subsets of $I_{i,j}$.

A working example is studied in Yan et al. [209], which computes the probability of inferring the user's typical paths (e.g., going to coffee shop, grocery, outdoors, etc.) only by exploiting the steps per minute computed from a fitness IoT pedometer. It has been reported that as long as the threshold value, ϵ, (denoting the Euclidean distance between the steps-tacked sequence and the path query sequence) varies, the user path could be inferred with at least approximately 50% of accuracy, thus, $P(userbehavior|pedometerdata) \geq 0.5$. Another example can be found in [31] where inferences from different smartphone sensor data are measured.

A related work on time series from Erdogdu et al. [50] computes the risks associated to general time series data using stochastic approaches.

AID-S Recommendation Strategies

As presented in Table 3.1, another main function of AID-S is its capability to recommend optimal solutions to ensure the privacy of the user. In this framework we propose two main strategies, either to recommend optimal privacy setting or transform the data.

- **Recommending the optimal privacy setting**

An optimal privacy setting is defined with respect to the set of data required by the third party in the Policy Statement. It is the set of personal data that represents the optimal balance between minimizing the risk of inferring personal data and maximizing the number of data to be shared with the third party. Maximizing the number of shared data is aimed to maximize the utility of the service provided by the third party.

From the user's point of view, AID-S can recommend which personal data item should not to be shared since it heavily increases the inference risk of another or a set of other personal data. Conversely, it can also recommend which data can be shared since it is not heavily correlated with any inference risk.

Depending on the specific AID-S implementation, the recommendation could be provided directly to the user, through the PDM's dialog manager, or it could be preceded by an attempt of automatic negotiation with the third party, aimed to balance the privacy of the user data and the utility of the third party service.

This recommendation strategy is efficient since data processing techniques (e.g., aggregation, transformation and obfuscation) will not be used. The whole process for privacy protection is based on the third party's Policy Statement and on the user's privacy preferences. For example, if the user does not want his location to be inferred by the third party, the privacy preference for this data item is set to P_{max}. In this work, this AID-S recommendation will be the use case during the realization of this framework.

Suppose a third party (e.g., fitness tracker) asks for the accelerometer data to PDM, AID-S is able to check all the correlations among data through $I_{i,j}$ and concludes that the accelerometer data can be shared since the probability of inferring

the user location does not reach the privacy threshold for location (equal to P_{max} in the example) for any combination with the accelerometer data.

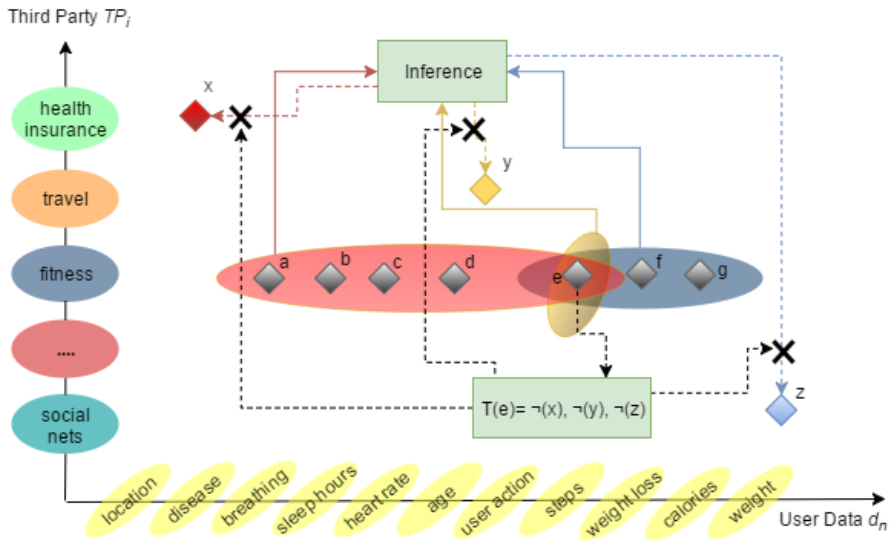


FIGURE 3.3: The effect of the transforming T(e) which removes the possibility of inferring of x, y, z.

To show the AID-S recommendation, Figure 3.3 shows a hypothetical situation in the Fitness IoT scenario. After the inference computation, AID-S finds that there is a high risk of inference for data x, given data (a,b,c,d,e). To reduce this risk, AID-S recommends to not share data e as it will significantly decrease the inference of x. In this scenario, this is the optimum recommendation.

- **Recommending data transformation**

The optimum recommendation might not work all the time, as both users and third parties might really need this data to be shared. Therefore, a suboptimal solution would be to perturb the data. The example in the figure illustrates the effect of applying a transformation T on data item e (using transformation techniques discussed in Chapter 2.1.2). Given that (a, b, c, d, e) are correlated with x and (e, f, g) are correlated with y, by transforming e, we obtain that two sets of correlations are broken, preventing the inference of x, y and z.

For AID-S, the transformation has to balance two requirements, which are the privacy of the user data and the utility of the third party service. This critical trade-off is the key for both parties to conclude an agreement. In the example above, one perturbation is able to reduce the risk of inference of three data items.

This concept is well defined in a study from Cai et al. [27] where they propose a method (i.e., collective method) to protect the user from possible inferences of third parties in social networks. The algorithm works as follows:

1. if PDAs ∩ UDAs = 0, then
2. remove PDAs
3. else, remove PDAs - core; and
4. perturbing core

In their model, an attribute (which is equivalent to a user data item, a_i, in this study) can be classified according to two criteria, either privacy or utility dependent. Privacy-Dependent Attribute (PDA) is when a user data is set or estimated as private, while Utility-Dependent Attribute (UDA) is when a user data is requested by the third party to provide its service. The first step of the algorithm checks if there are no user data items that belong to both PDAs and UDAs. If none, then PDAs will be removed and only UDAs will be released to the third party. If there is, then this user data (termed as core) must be perturbed before being released to the third party.

To perform perturbation, they use a classical method of substituting the core with a more abstract or generalized annotation. This perturbation technique has several levels of hierarchy and is called Generic Attribute Hierarchy (GAH). For example, instead of releasing the specific user attribute on the category "Favorite music", it can be perturbed as slow music. Different approaches can be found in the literature to manage this task (e.g., [5, 27]).

Different data transformation/perturbation methods can be found in literature as discussed, and will not be the focus of this work. Literature approaches can be utilized in this recommendation strategy, which is the suboptimal solution if the recommendation of the optimal settings will not be possible.

AID-S Inference risk discovery and recommendation using optimum privacy settings are elaborated in Chapter 5, which answers RQ1.

Chapter 4

Fitness IoT Scenario

This chapter discusses the fitness IoT domain, which is the chosen IoT domain to apply our framework. It includes discussion on today's scenario regarding the different permissions asked by fitness apps, the complex information-sharing model of these apps, the modeling of the IoT fitness data and the prototype used for simulation for our data collection.

It should be noted that our approach will be applied, but is not necessary limited to, the fitness IoT domain. Fitness IoT is a growing area inside the IoT wearable systems, i.e., IoWT, which currently holds the most number of user personal data as it monitors user's round-the-clock user activities (24/7).

4.1 Today's Scenario

Fitness trackers are becoming popular today and are predicted to continually grow in the coming years [38]. This also means that a lot of manufacturers and service providers are in competition with different device types, features, software, systems, etc. that make users unsure about which product to purchase. Also, besides the apps provided by the producers of the fitness trackers, there are other third party apps that are compatible with the trackers. They also provide services by accessing, if authorized by the user, the sensed data collected by the tracker. In this section, we briefly analyze a set of fitness devices currently in the market showing how they get into the user's daily life.

TABLE 4.1: Fitness Trackers Comparison

Brand	Device	Sensors	Main processed data
Fitbit[1]	Surge	GPS, 3-axis accelerometers, 3-axis gyroscope, digital compass, optical heart rate monitor, altimeter, ambient light sensor, vibration motor	activity and exercise, heart rate, elevation, steps, distance, floor, calories, sleep, profile (e.g., weight, height), nutrition (e.g., food and water logs)
Garmin[2]	Vivosmart HR+	barometric altimeter, accelerometer, heart rate monitor, GPS-enabled	single, multi-activities, multi-sport with heart rate, steps, distance, calories, flights of stairs, intensity minutes, heart rate, run/walk interval, virtual pacer (helps to achieve goals)
Jawbone[3]	UP4	3-axis accelerometer, bioimpedance sensor, accelerometer, motion sensor	band events, body, heart, goals, meal, mood, moves, sleeps, trends, workouts, user information, calories
Misfit[4]	Shine2	3-axis accelerometer, magnetometer, capacitive touch sensor	steps, distance, calories, and light and restful sleep

4.1.1 Fitness trackers

Fitness trackers, as shown in Table 4.1, assist users in keeping fit by tracking their activities and workouts (e.g., daily steps taken, floors climbed, light/moderate/intense activities, cycling/swimming time, resting time, etc.), computing the related fitness measures (e.g., calories burned, quality of sleep, cardiovascular workouts, calories intake, etc.) and suggesting them their suitable personalized programs. Also, these trackers have features to enable fitness challenges with friends for increasing the user's motivation.

Table 4.1 shows the comparative features of the most popular fitness trackers in the market. Their features are described in terms of the data that they collect and the sensors that they use. Only the trackers with the most features and sensors are shown regardless of their price and application (e.g., specific for swimming, running, etc.) since this study will focus on the information that the trackers can produce.

In summary, fitness devices track the user's precise movements, location, sleep dynamics, health information, preferences, etc. As of today, these devices have many details of personal information and monitor the user 24/7 with persistency of at least per-minute of accuracy. Allowing and authorizing third party apps to access and process one's personal data represents a potential risk for users.

4.1.2 Android permissions

A third party app specifies in its policy statement the list of user information it needs in order to provide its service. These policy statements in an Android device are in Android Permissions (precisely in the Android Manifest XML file of the application). Since Google's Android is an open source operating system based on the Linux kernel, it inherits from Linux the file permissions mechanism by which files are owned by specific users with read/write or execute (*rwx*) permissions. A file has a defined *rwx* also for a group IDs.

Furthermore, because of the Android application-permission mechanism, each application has an application specific user with a unique user ID (UID), which owns its applications files, while the system files are owned by system or root user. Android automatically grants permission to an app to use some resources needed during installation. These permissions are known as the normal permissions and they are not asked to the user. Any additional permission[5] such as accessing fine location (ACCESS_ FINE_ LOCATION) (e.g., GPS), recording audio (RECORD_ AUDIO) or sending SMS messages (SEND_ SMS), must be explicitly requested and authorized by the user installing the application. These permissions are defined as dangerous permissions[6] by Android since they may contain user information. In Table 4.2 we report the permission groups and their respective dangerous permissions. Our study concerns these permissions. It must be noticed that for Android 5.1 and lower versions, when installing an app, the user must grant the whole block of permissions requested by the app. Differently, for Android 6.0 and higher the permission can be granted on runtime.

[5] https://developer.android.com/reference/android/Manifest.permission.html
[6] https://developer.android.com/guide/topics/manifest/permission-element.html

TABLE 4.2: Android permissions groups and the dangerous permissions per group.

Permission Group	Dangerous Permissions
Calendar	READ_CALENDAR WRITE_CALENDAR
Camera	CAMERA
Contacts	READ_CONTACTS WRITE_CONTACTS GET_ACCOUNTS
Location	ACCESS_FINE_LOCATION ACCESS_COARSE_LOCATION
Microphone	RECORD_AUDIO
Phone	READ_PHONE_STATE CALL_PHONE READ_CALL_LOG WRITE_CALL_LOG ADD_VOICEMAIL USE_SIP PROCESS_OUTGOING_CALLS
Sensors	BODY_SENSORS
SMS	SEND_SMS RECEIVE_SMS READ_SMS RECEIVE_WAP_PUSH RECEIVE_MMS
Storage	READ_EXTERNAL_STORAGE WRITE_EXTERNAL_STORAGE

4.1.3 Case study

In this scenario, we describe how typical users utilize and perceive fitness trackers and the related applications. We assume a user is equipped with an Android 5.1 (AOI model) smart phone and decides to buy a Fitbit tracker.

Upon buying the device, the user is required to create a Fitbit account on the Fitbit website to access the data processed by Fitbit. Moreover, to connect the device with the smart phone, the user is required to install the Fitbit app. When installing, the **Fitbit app**[7] asks for the following user permissions to access smartphone resources:

- Identity
- Contacts
- Location

[7] https://play.google.com/store/apps/details?id=com.fitbit.FitbitMobile&hl=en

- SMS
- Photos/Media/Files
- Camera
- Bluetooth connection information
- Device ID and call information

It should be noted that the above permissions do not correspond exactly to the group names in Table 4.2, which are for development purposes.

Given that the user is equipped with Android AOI model, his option is either to accept or reject all the permissions asked by the Fitbit app. Users having the latest versions of Android can select which permission to grant. Additionally, during the registration, Fitbit app requires the following further personal information as mandatory.

- Name
- Gender
- Height
- Weight
- Birthday

After the user has registered and installed the Fitbit app, he/she can start exercising and enjoying the benefits of the Fitbit services. The user's data can be analyzed and viewed in the Fitbit dashboard and can also be exported in csv/xls format.

Third party sharing

Aside from the main Fitbit App, there are a lot of third party apps that are compatible with the Fitbit devices and can be authenticated to the Fitbit server. This trend is becoming popular today among different third parties. These further apps can improve and provide new features to support the user's fitness routine/goals. For a working example, the **Lose IT! app**[8] will be used. It helps users to lose weight by

[8] https://play.google.com/store/apps/details?id=com.fitnow.loseit&hl=en

TABLE 4.3: Specific permissions requested by the Fitbit and Lose IT! app to Android OS

	Fitbit app	Lose IT! app
Main	INTERNET	INTERNET
	ACCESS NETWORK STATE	ACCESS NETWORK STATE
	WRITE EXTERNAL STORAGE	WRITE EXTERNAL STORAGE
	CAMERA	CAMERA
	WAKE LOCK	WAKE LOCK
	ACCESS FINE LOCATION	ACCESS FINE LOCATION
	GET ACCOUNTS	GET ACCOUNTS
	BLUETOOTH	BLUETOOTH
	BLUETOOTH ADMIN	
	READ CONTACTS	
	NFC	
	READ EXTERNAL STORAGE	
	RECEIVE BOOT COMPLETED	
Other	com.fitbit.FitbitMobile.permission. C2D MESSAGE	com.android.vending.BILLING
	com.google.android.providers.gsf. permission.READ GSERVICES	com.fitnow.loseit.permission. C2D MESSAGE
	com.google.android.c2dm. permission.RECEIVE	com.google.android.c2dm. permission.RECEIVE
		com.microsoft.band.service.access. BIND BAND SERVICE

estimating the calorie intake of the user by taking a picture of the user's meal. When installing the Lose IT! app, it asks for the following permissions on the smartphone resources as mandatory:

- In-app Purchase
- Identity
- Location
- Photos/Media/Files
- Camera

Table 4.3 summarizes the Android OS permissions requested by the Fitbit app and the Lose IT! app. After installing the Lose IT! app, the user is further asked for the following mandatory information, which is similar to the set of in-app permissions asked by Fitbit:

- Birthday

- Height
- Weight
- Gender

In order to enable Lose IT! to process the data collected by the Fitbit tracker, the user is redirected to the Fitbit permission page which asks the user for the permission to allow or deny Lose IT! to the following Fitbit data:

- Sleep
- Food and water logs
- Activity and exercise
- Weight

Given the similar aims of the two apps, some of the data are used by both Fitbit and Lose IT! such as weight and food logs since both are used for computing the absorbed calories. In our running example, absorbed calories will be referred to Lose IT! and the weight will be referred to Fitbit.

When a user accepts to share, for instance *activity and exercise*, he/she may be unaware of the effective personal data both sensed and estimated (as shown in Table 4.4) that he is disclosing with different intensity of activity (e.g., sedentary, very, fairly, and lightly active). Thus, due to the lack of adequate explanation and definition in the permission warnings [55, 114], users may not have understanding of what sharing implies (in our case *activity and exercise*), and as a result, they could take unsafe decisions.

4.1.4 Third party service model

The sharing of user information with third parties is shown in Figure 4.1. We created a simplified model of the complex scenario of third party service chains.

Third Party (TP) apps, such as Lose IT!, Strava and RunKeeper in the figure, require an integration with Fitbit to comply with their API Terms of Service and must be registered at *dev.fitbit.com*, specifying the set of data they will use in their service among those exposed by the Fitbit API.

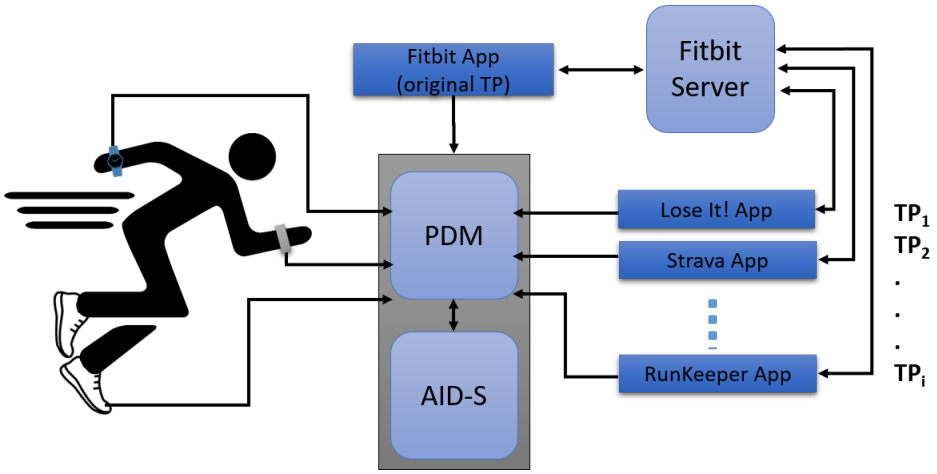


FIGURE 4.1: Architecture combining the service model of today's fitness trackers with the PerNANDO framework.

A user who wants to use the feature of integration between the third party app and Fitbit will be redirected to Fitbit's authorization page in which, through the OAuth 2.0 protocol, he/she has to allow/deny access to Fitbit data required by the third party app. For the case of the Lose IT! app, the data required from Fitbit are sleep, food and water logs, activity and exercise, and weight. Table 4.4 reports a subset of data exposed by the Fitbit API. It is worth noting that each category of data (i.e., Scope in the Table 4.4) can include a subset of personal data.

Moreover, the current fitness sharing model is more complex since the accessing TP app itself can provide to Fitbit the user data it collects or elaborates (i.e., two-way access). In other terms, some sets of user data are synchronized to Fitbit given the user permission. For instance, Lose IT! reads the activity and sleep data from the Fitbit server, writes the food log on the Fitbit server and reads the calories burned from the Fitbit server.

Definition 1 (User Data Sets)**.** *We define S, A and $F \in D$ as the sets of possible user information asked by an app in the installation process, respectively, as the following:*

S: smartphone's data required by the app through the operating system's permission mechanism (smartphone permission set),

A: personal data required by the Application (in-application set) for the service it provides,

F: fitness data shared by the original service, Fitbit in our example, (inter-application set),

as subsets of the set of all the possible user information, D.

TABLE 4.4: Subset of Fitbit data provided by the Fitbit API. Each category (scope) includes further data items.

Scope	Data Subsets (*F* data sets from Fitbit)
Activity	- activities (calories, distance, duration, steps): - activity logging - based on the activities in a catalog (running, bicycling, walking, etc.) - activities_calories -includes tracked activity + manually logged activities + BMR - activities_caloriesBMR - only BMR calories - activities_activityCalories - the number of calories burned during the day when the user is active above sedentary level, including BMR - tracker_activityCalories - similar to aforementioned by its calculated using only mbox tracker data (i.e., without BMR) - tracker_calories - calories burned inclusive of BMR according to movement captured - goals (caloriesOut, distance, floors, steps) - summary (activityCalories, caloriesBMR, caloriesOut, distance, elevation, fairly/lightly/very ActiveMinutes, floors, sedentaryMinutes, marginalCalories, steps) - steps, distance, floors, elevation
Heartrate	- restingHeartRate - heartRateZones max and min - heartRateZones name (Out of Range; Fat Burn; Peak; Cardio) - heartRate zones minutes - heartRateZones caloriesOut
Profile	- weight and height - encodedId, memberSince, etc. - averageDailySteps - strideLengthWalking and strideLengthRunning - aboutMe, avatar, city, country, dateOfBirth, fullName, gender, etc.
Sleep	- timeInBed - minutesAsleep and minutesAwake - awakeCount, awakingDuration - restlessCount, restlessDuration - startTime - minutesToFallAsleep - efficiency - minutesAfterWakeup - goals data - summary data
Nutrition	- food_logs (foodName, mealTypeId (breakfast,morning snack, lunch, etc.), amount, date, favorite, brandName, calories) - water_logs (amount, date and unit) - food_goal (foodPlan, calories) - water_goal - nutritionalValues associated to a specific food divided in Common, Vitamins and Dietary Minerals (calories, carbs, fat, fiber, protein, sodium) - food_locales - specifies the country
...	...

Definition 2 (Third Party Data Sets). *A Third Party Fitness App, TP_i, requires permission for user information sets composed of $S_i \in S$, $A_i \in A$, $F_i \in F$ where, each smartphone user data $s_x \in S_i (1 \leq x \leq |S_i|)$, in-application user data $a_y \in A_i (1 \leq y \leq |A_i|)$, and inter-application user data $f_z \in F_i (1 \leq z \leq |F_i|)$.*

Thus, a permitted TP has all these pieces of information, i.e., S, A, F sets, for a single user as well as of the other users who granted permission to access their data. For a user, it is hard to maintain track of all his/her shared information especially as the number of third party apps installed on his/her device tends to increase. Moreover, it is hard to figure out the data that can be inferred not just from his/her own data, but from the entire TP dataset.

4.2 Fitness IoT Data Model

Previous section provides an overview of the current Fitness IoT scenario. The complexity clearly calls for a privacy manager that can ease the user's permission setting. For this to be realized, this section aims to derive the data permission model for the IoT fitness domain, which is one of the task of PDM, as discussed in Chapter 3. More importantly, the data permission model is made sure to be GDPR-compliant.

Using the same scenario in the previous section, other major fitness trackers app that can provide data to other third party fitness apps are also considered for the data model.

Table 4.5 shows a comparison of the data collection practices of four existing fitness trackers. We selected these trackers based on their popularity[9] and the maturity of their software solutions (i.e., working Web API for integration). Our list is comprised of Fitbit[10], Garmin[11], Jawbone[12], and Misfit[13].

As shown, the requested data can be categorized into three sets. The first set of requested data involves the smartphone permissions, which are requested during the installation or the first use of the app. We define this set as the *S set*. The next set

[9] https://www.wareable.com/fitness-trackers/the-best-fitness-tracker
[10] https://play.google.com/store/apps/details?id=com.fitbit.FitbitMobile
[11] https://play.google.com/store/apps/details?id=com.garmin.android.apps.connectmobile
[12] https://play.google.com/store/apps/details?id=com.jawbone.up
[13] https://play.google.com/store/apps/details?id=com.misfitwearables.prometheus

TABLE 4.5: Comparison of permissions asked by Fitness Trackers and the fitness IoT Data Model used for this study.

	Fitbit	Garmin	Jawbone	Misfit	Our Data Model
(S Set) Phone	• Bluetooth • Camera • Contacts • Device & Call • Identity • Location • Photos/Media • SMS • Storage	• Bluetooth • Calendar • Camera • Contacts • Device & Call • Identity • Location • Phone • Photos/Media • SMS • Storage • Wifi Inf.	• Bluetooth • Camera • Contacts • Device & Call • Identity • Location • Microphone • Phone • Photos/Media • SMS • Storage	• Bluetooth • Camera • Contacts • Device & Call • Identity • Location • Phone • Photos/Media • SMS • Storage	• Bluetooth • Camera • Contacts • Identity • Location • Media & Music • Mobile Data • Motion & Fitness • Phone • Photos • SMS • Storage
(A set) In-app Requests	• Birth date • Gender • Height • Name (First) • Name (Last) • Weight	• Birth date • Gender • Height • Name (Display) • Name (Full) • Weight	• Birth date • Gender • Height • Name (First) • Name (Last) • Weight	• Birth date • Gender • Height • Name (Full) • Occupation (optional) • Weight	• Birth date • Gender • Height • Name (First) • Name (Last) • Weight
(F Set) Fitness Data	• Activity & Exercise – activity minutes – calories activity – distance – elevation – floors – steps • Devices & Settings • Food & Water Logs • Friends • Heartrate • Location & GPS • Profile • Sleep • Weight	• Full Fitness data • Location • Sync Device	• Basic Info • Extended Info • Heartrate • Meals • Moves • Sleep • Friends list	• Device • Goal • Profile • Session • Sleep • Summary – activity calories – calories – distance – steps	Activity & Exercise • activity minutes • calories activity • distance • elevation • floors • steps • Devices & Settings • Food & Water Logs • Friends • Heartrate • Location & GPS • Profile • Sleep • Weight
(G Set) GDPR					Entity Types • Fitness/Health apps • FP (corp.) • FP (gov't.) • SN (friends) • SN (public) • other apps Purposes • convenience • commercial • health • safety • social • Frequency • Retention

of requested 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*. More importantly, the data 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 4.5 is the most common set of data items collected by the four trackers, taking into account the different mobile operating systems. Moreover, it includes also the *G set* which concerns the GDPR requirements that will bee explained in Section 4.2.4. All the data items in the last column form the Fitness IoT Data Model for this study.

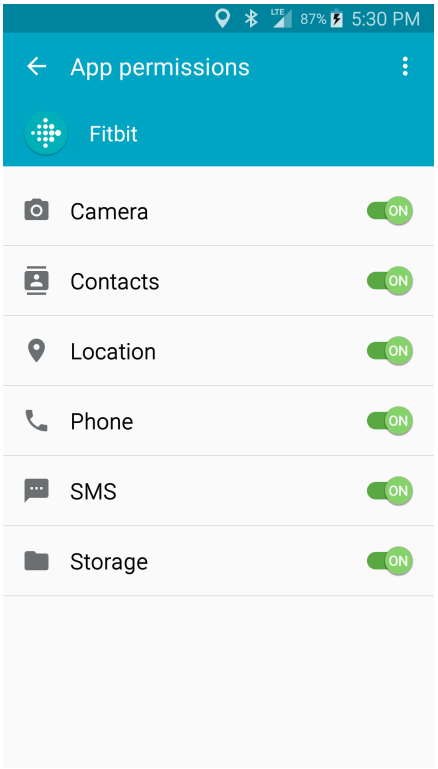
4.2.1 The S set (smartphone permissions)

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 AOI in version 5.9 and below to AOFU in version 6.0 and above. While Table 4.5 considers the Android AOI permissions requested by various fitness apps, Figures 4.2a and 4.2b show the Fitbit's permissions for Android AOFU and iOS for comparison.

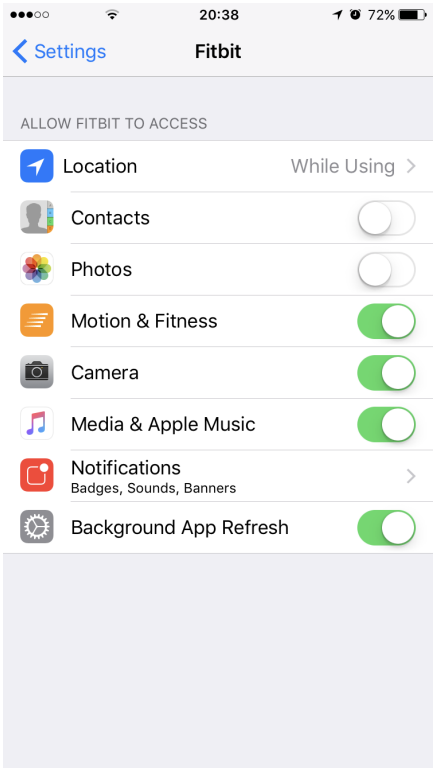
The final data model for the S set is composed of the mostly commonly used set of the different mobile OS permissions requested by the fitness apps in Table 4.5. The *background App* and *Notifications* iOS phone permissions are not taken into account since these permissions are not relevant for third-party data access. *Other* permissions in Android AOI, which are not general, are also not taken into account. The permission for *Photos/Media/Files* in Android AOI was divided into *Photos* and *Media & Music* to take into account the granularity of iOS permissions. We finally have a total of 12 permissions for the S Set.

4.2.2 The A set (in-app requests)

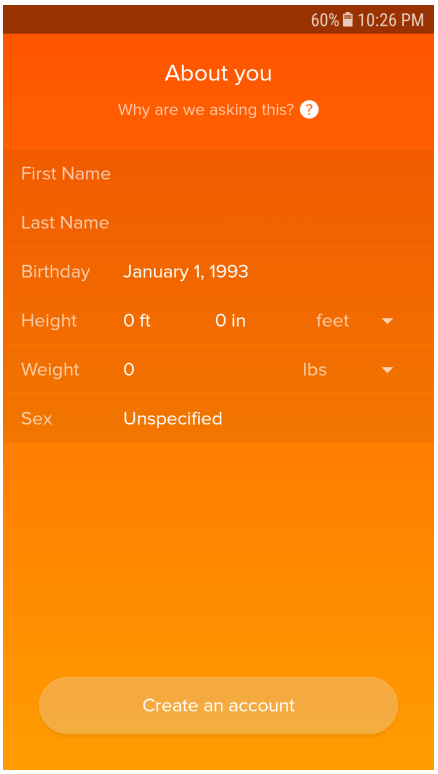
In addition to the smartphone permissions that fitness tracking apps ask, they also gather information inside their application, e.g., as part of the sign-up process for their online services. This data usually includes the user's *First Name* and *Last Name*,



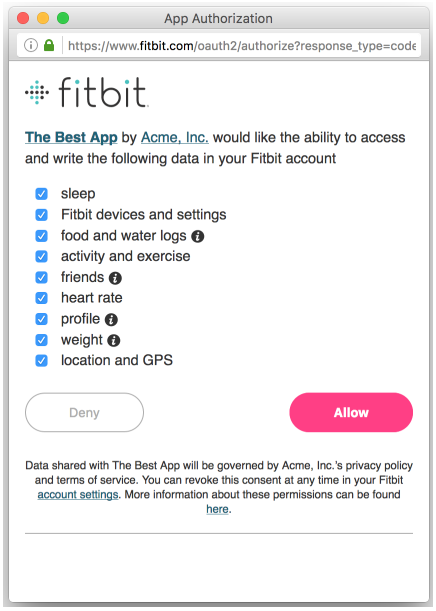
(A) S set (Android 6.0+ Model).



(B) S set (iOS Model).



(C) A set.



(D) F set.

FIGURE 4.2: Examples of Fitbit permission requests.

Birth Date, *Gender*, *Height* and *Weight*. Note that these data are mandatory for all fitness trackers in Table 4.5; the only optional piece of information is Misfit's request for the user's *Occupation*. Figure 4.2c shows the *A set* for the Fitbit app (other apps are similar), as reported in Table 4.5. A total of 6 permissions are considered as part of the A set.

4.2.3 The F set (fitness data)

The F set contains the data fitness trackers collect while the user is using the device. Some of this data is 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 this data is "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 this data with other TPs in exchange for additional fitness or health services the user can benefit from. Table 4.5 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 Figure 4.2d[14]. Additionally, these settings can be revisited in their web app[15], where users have the option to revoke the granted access. The other apps in Table 4.5 also give users control, but only give users 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 [181]. 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

[14] https://dev.fitbit.com/build/reference/web-api/oauth2/
[15] https://community.fitbit.com/t5/Flex-2/How-do-I-revoke-access/td-p/2701359

his/her fitness tracking data. In this study, instead of taking into account individual TPs, we use the PPIoT *EntityType*, which will be discussed in Chapter 6.1, to investigate which group category of TP apps (namely "who") the user prefers to share with. This parameter has been shown to be important in determining users' privacy settings [108]. Since Entity Types are intimately related to GDPR-based requirements, these permissions are included in the G set.

4.2.4 G Set (GDPR-based permissions)

The G set includes permissions that are based on GDPR requirements. We report the exact terms used in PPIoT Ontology 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*.

Given today's data sharing model, fitness data can be shared to different fourth party apps, which are mostly composed of fitness/health apps[16] or social media apps[17]. For this reason app categories are divided into fitness/health apps, social media apps, and other apps to take into account the minority apps (e.g., game apps that access fitness data). Additionally, work-related fitness programs are also included given the increasing adaption of employers to promote wellness of their employees [52]. Fitness programs allow to reduce stress in work environments. Mainly, these entities can be categorized into corporate fitness programs or government fitness programs [22] (e.g., for military).

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

[16] https://www.fitbit.com/partnership
[17] sharing to social media: https://help.fitbit.com/articles/en_US/Help_article/2106

We did not include the *hasMethod* property since it involves technical background, as stated in Section 2.6, which may not be known to the users. For simplicity, we assume that the TPs' *hasMethod* data access are *encrypted*.

4.2.5 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, includes 45 permissions in total. Moreover, if the user wants 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 very clear manner, 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 semi-automated interactive *privacy recommendations* to partially relieve users' burden of setting each of these individual permissions and meanwhile maintain the control on privacy preferences.

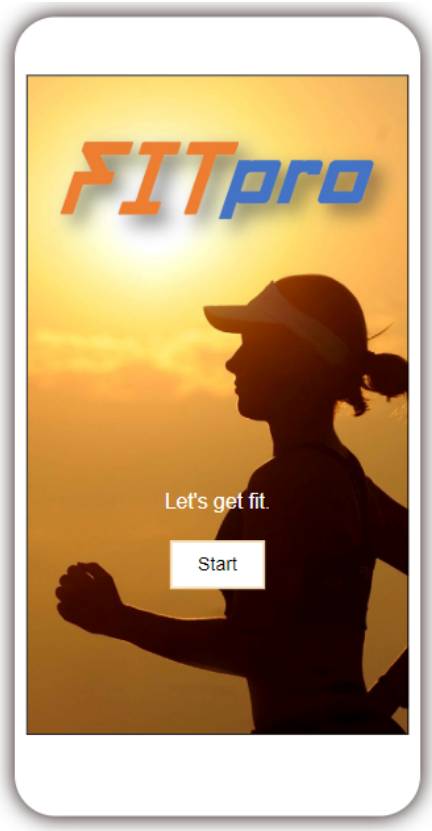
4.3 Data Collection

This section provides the details of our data collection. Mainly there are two separate datasets for the AID-S and PDM. For the AID-S, the data needed are real fitness data from users. The real content will be studied to see which of data have correlation to the others. On the other hand, PDM does not require real content of the data but the user's perception on sharing such data. Thus, it is more difficult to obtain such data.

4.3.1 AID-S Dataset

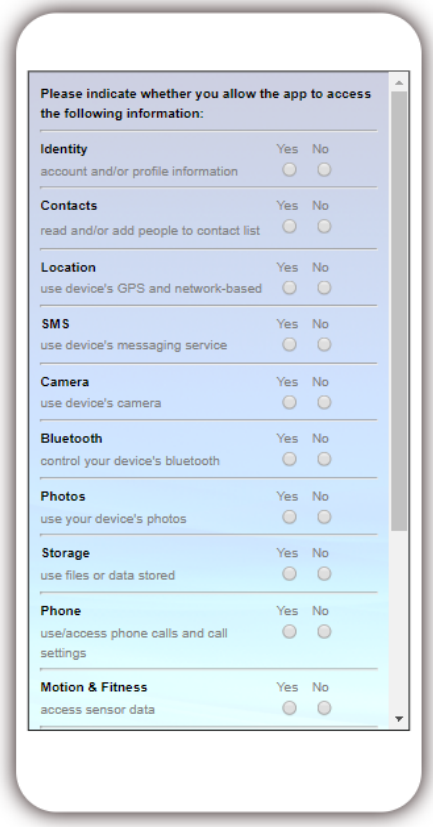
For the AID-S inference analysis, we used the public data set of Fitbit users from the Open Humans Foundation[18] and the crowd-sourced Fitbit dataset generated from

[18] https://www.openhumans.org/

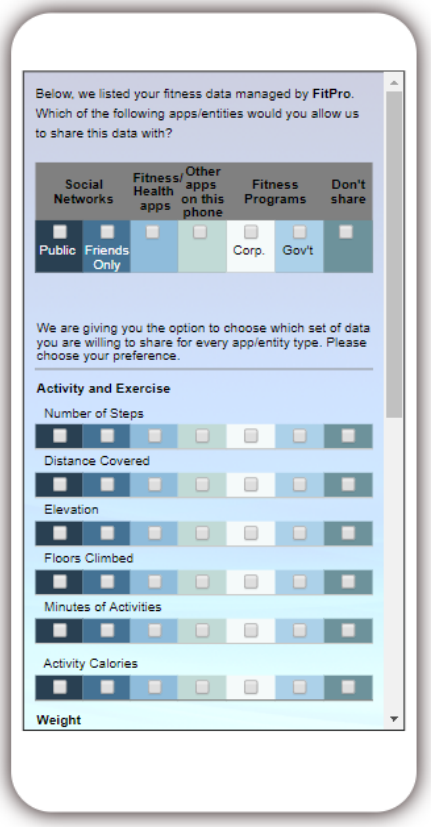


(A) FitPro App

(B) In-app requests (A set)

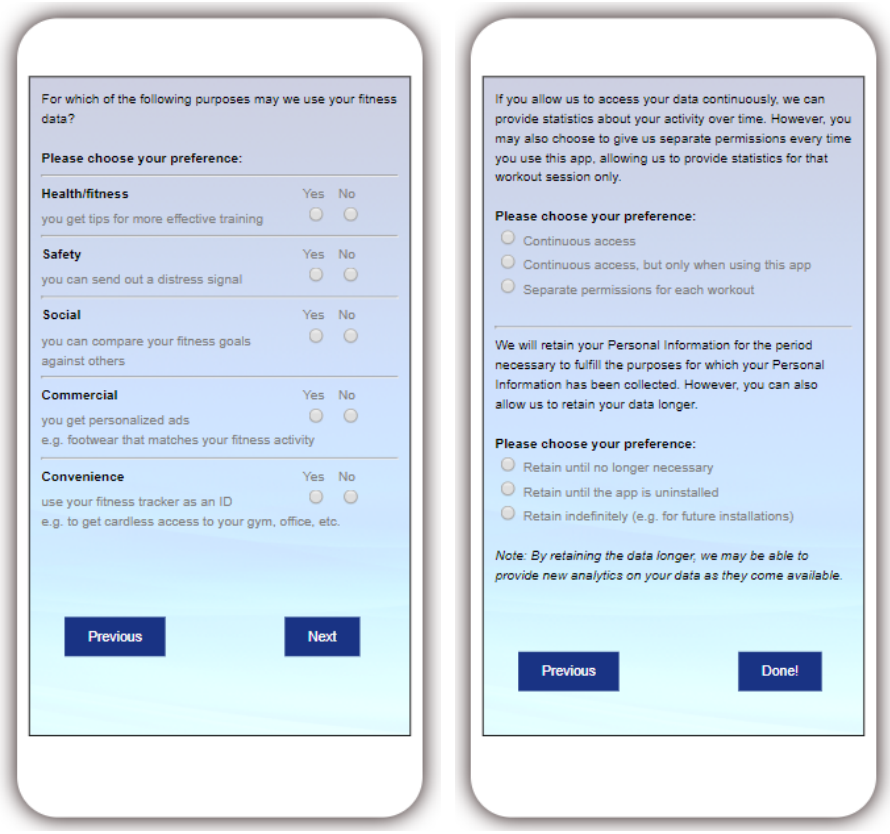


(C) Phone requests (S set)



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

FIGURE 4.3: Fitpro (app prototype).



(E) Purpose of collection (G set) (F) Frequency & retention (G set)

FIGURE 4.3: Fitpro (app prototype) (cont.)

the respondents of a distributed survey via Amazon Mechanical Turk [19] by Furberg et al. We used a total of 19,817 samples from 49 users. As of the time of writing, the first and the second dataset store time series data of 14 and 35 Fitbit users respectively.

The datasets consist of time series data regarding the user's number of steps taken, distance traveled, minutes of activity, floors taken, elevation, activity calories, weight, minutes of sleep, and heartrate information. For some users there is no information about some features, i.e., elevation and floors, which are available only for trackers with an altimeter, and heart rate information for trackers that do not have the heart rate monitor. We considered only the daily data series from the tracker device excluding the manually entered activity values. For example, we considered the resource path activities/tracker/steps (data acquired exclusively from the device) and not the path activities/steps (which includes also the manually entered

[19] https://zenodo.org/record/53894#.W-76oehKjIU

values when an activity is logged by specifying distance and time) [20].

The datasets were preprocessed, noise was removed (e.g., empty sets when users are not using the tracker), and ranges were normalized resulting in a total of 6,229 clean samples. In machine learning approach, required training sample size varies depending on the problem complexity (e.g, no. of classes and features) and classifier complexity. Following the rule of thumb that a sample size must have at least 50 for small sets features (e.g., 5-10 features) [152] (e.g., the same for bayesian approach in [211]), we conclude that our sample size is enough for this study.

4.3.2 PDM Dataset

The data collection for PDM is more challenging. To collect a sample of fitness IoT permission settings for the PDM User Profiling and Recommendation, we created a mock app, which will be discussed further below, that follows the PPIoT-based and GDPR-compliant privacy model. Then, we recruited 310 participants through AMT to simulate our mock app. After data preprocessing, we utilized a total of 295 samples.

We asked people to only participate if they were active Fitbit users[21], and checked this requirement by asking participants to enter the initial and last few digits of their Fitibit serial number. The participants consisted of 34.2% males and 65.8% 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 Data Model for Fitness IoT that we defined in Chapter 4.2, as shown in the last column of Table 4.5. Each participant used this prototype, followed by a questionnaire.

[20] https://dev.fitbit.com/docs/activity/#activity-time-series

[21] 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.

4.3.3 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[22]. As shown in Figure 4.3, the permission setting interface of *FitPro* consists of the following parts:

- Figure 4.3b 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 discussed in Section 5.1) 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 4.3c shows the permissions UI for the *S set*—the permissions the TP needs to attain 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 4.3d 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 request about "what data" can be accessed by "who". As discussed in Section 4.2.3, 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.

- Figures 4.3e and 4.3f show the permissions UI screens for the *G set* 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.

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

4.3.4 Questionnaire

After using the prototype, we asked participants to fill out a questionnaire[23]. The goal was to investigate if certain user traits, some of them already studied in the literature, have relations with the users' privacy behaviors collected through FitPro. Specifically we aimed to measure the users' privacy-related attitudes (trust, privacy 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 A.1 in the Appendix.

Privacy Attitude

Our privacy attitude questions consist of 5 topics that were used to study different attitudes. Questions on user's trust in app provider were derived from [96] and [175]. Questions on general privacy concerns are based on [123] which was originally based on [168]. User's perception of surveillance, intrusion, and secondary use of personal information are from [206]. Perceived intrusion and secondary use of personal information questions are originally from earlier work by [205] and from [168], respectively. These user attitudes are used extensively in the privacy literature and are proven to have significant effects on users' privacy behaviors.

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 [150]. Moreover, in the field of privacy users' decisions tend to depend on the risks and benefits of disclosure, as we found in previous work [92]. Following this approach, we take the negotiability of users' privacy settings into account in this study and we measure it as event-based change of preference: we ask users' 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).

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

Social Behavior

Sociability (i.e., the ability to interact) also is a factor to predict links of users [166] not necessarily with 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.

Exercise Tendencies

These questions are grouped in two topics: exercise attitudes and healthy living expertise. The former are fully self-developed. Our aim is to investigate if users' exercise attitude (e.g., the intensity of exercise, type of exercise, how healthy users are, how important exercise is to the users, the reason for exercising) influence their tendency to allow fitness apps to collect and share their data. The healthy living expertise questions are taken from [91], and measure how knowledgeable people 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.

User Demographics

User demographics such as gender, age, location, and education are often used to improve recommendation accuracy [150]. In our questionnaire, we introduce this category to investigate if there is an association between users' privacy settings and their demographic attributes, as resulted in previous studies (cf. [96]).

Chapter 5

AID-S Inference Protection

The need for a service that enables the user to take informed decisions is relevant. This is due to the fact that not only the number of third party applications increases but also the number of heterogeneous IoT devices. The Adaptive Inference Discovery Service (AID-S) of the PerNANDO Framework, as shown in Figure 4.1, is proposed to address such issue [180]. The core concept of AID-S is to provide a service that aims to compute the risk of possible inference of private information.

As discussed in Chapter 3.2.2, the Inference matrix, $I_{j,k}(TP_i)$, expresses the possibility for a TP to infer personal data a_j which the user denied to share when he/she installed the app or while using the app. This kind of inference is often called in the literature as *attribute disclosure* [173] and means inferring the value of private data based on the integration and correlation of non-private data [5]. AID-S aims to estimate this risk matrix for each TP_i.

5.1 Inference Graphical Model

We use a graphical model to depict the network of correlations among the user data. This graph will be used to identify the possible inferences using Bayesian Networks.

5.1.1 Bayesian networks

Bayesian Networks (BNs) are graphical representations of knowledge. They are Directed Acyclic Graphs (DAGs) which are composed of nodes and arcs that represent the variables and their corresponding probabilistic relations, respectively [144].

These networks are used in broad range of applications such as from Business Intelligence (e.g., Online Analytical Processing (OLAP)), medical (e.g., service performance analysis [1] and breast cancer prognosis and epidemiology [76]), biological sciences (e.g., gene expression analysis by [58]) and many more [160]. BNs are efficient in handling applications filled with deep uncertainties by combining prior knowledge with observational data in order to infer and model causal relations [160].

There are two general rules that must be respected in order for a network to be a BN:

- it should be a DAG, which means that by following the directed arcs, it should not be possible to return to the parent node [101],
- the Markov condition must be satisfied which states that each variable in the DAG must be conditionally independent of its non-descendant, given the set of all its parents [133].

Therefore, for a DAG ($G(V, A)$ where V and A are the set of random variables and arcs, respectively), the factorization of the joint probability distribution of the set V, considering the Markov Condition, can be computed as:

$$P(X_1, X_2, ..., X_v) = \prod_{i=1}^{v} P(X_i|Pa_i) \tag{5.1}$$

where $X_i \in V = \{X_1, X_2, ..., X_v\}$ is the i^{th} random variable in the Graph G, and Pa_i is the parents' nodes of the i^{th} random variable. Equation 5.1 shows the explicit need for the conditional probability between itself and its corresponding parent. Then, the chain rule will be used to compute the joint probability of all nodes.

To create a BN, the structure of the network (the topology given by the DAG) and the conditional probabilities of the parent/descendant nodes must be taken into account. Respectively, the topology and the conditional probabilities can be found through the use of Structural and Parameter Learning algorithms.

Structural learning

Structural Learning aims at finding the topology of the BN. These algorithms can be classified into three major groups, namely, constraint-based algorithms, score-based algorithms, and their combination known as hybrid algorithms.

Constraint-based algorithms are based on conditional independence statements (i.e., constraints). These algorithms learn the topology of the network through a conditional independence test (e.g., X^2 test) to create the edges among the variables and add the direction of the edges that satisfy the d-separation criterion [145].

Score-based algorithms assign a score that corresponds to the goodness of a BN structure with respect to the data set. The score is evaluated for all the possible candidates of a BN structure which is generally maximized through heuristic search techniques.

For a large number of variables, constraint-based algorithms are superior over the greedy search score-based algorithms. However, relying on d-separation statements always results in an incomplete directed network. Also, they are very prone to failures in the dependence tests [172]. For a smaller data set, score-based algorithms tends to be more efficient as it can search in the space all the possible structures and find the optimum. These factors must be taken into account in choosing the learning algorithms.

It is also possible to combine the two strategies, which are generally called hybrid algorithms [172]. In this case, the common approach is to create the undirected graph through the constraint-based algorithms while the directions of the arcs are learned by the score-based algorithm.

Using the *bnlearn*[1] package for the R[2] language enables us to implement: constraint-based structure learning algorithms (i.e., Grow-Shrink (GS), Incremental Association Markov Blanket (IAMB), Fast Incremental Association (Fast-IAMB), Interleaved Incremental Association (Inter-IAMB), Max-Min Parents and Children (MMPC), Semi-Interleaved Hiton-PC (SI-HITON-PC), CHOW-LIU, and ARACNE), score-based structure learning algorithms (i.e., Hill-Climbing (HC), Tabu Search (Tabu)) and hybrid

[1] http://www.bnlearn.com/
[2] https://www.r-project.org/

structure learning algorithms (i.e., Max-Min Hill Climbing (MMHC), General 2-Phase Restricted Maximization (RSMAX2)).

Parameter learning

The estimation of the prior and conditional probabilities related to the arcs have to be done after the topology is completed. Parameter learning aims to estimate the probabilistic relations of the BN structure.

bnlearn also comes with parameter learning that uses Maximum Likelihood parameter estimation for both discrete and continuous data sets and Bayesian Estimation for the discrete data sets [161].

5.2 Experimental Methodology

5.2.1 Objectives and description

Goals. In this section, we describe our methodology for experimentally applying the principles of the framework. Specifically, the aim of this section is to present a practical approach to implement the AID-S inference reasoning task (named *Inference risk computation* in Table 3.1) by modeling the dependencies among user data through a BN.

BN is used since our goal is to create a graphical inference network and extend the current network of inferences in the literature. Also, BN is a probabilistic approach used in modeling risks such as in [56], which is suitable for this study.

Steps. As it arises from the framework description, our approach does not perform any traffic analysis aimed to identify possible inference risks based on the content of the shared data. Our approach is *blind* with respect to the user data since we do not analyze nor store any data. The inference risk prediction is essentially based on the knowledge stored in the network in terms of dependencies among data.

Our approach is an advantage for the user since, as a third party, the framework does not need to know the content of the data, only its description. This is a safer for users compared to other approaches. However, our functionality is limited to computing inference risk, but that is all the AID-S service needs to achieve.

The use of BN makes it easier for the process of extending the graph with new personal data, given the availability of knowledge about their dependencies. Two approaches to acquire knowledge and extend the inference graph are the following:

- identifying the dependencies using a ground-truth dataset,
- exploiting the studies available in the literature and the correlation indexes provided in these studies.

We will use both the approaches and will validate them experimentally. The steps are as follows:

- Step 1: we will use the first approach to identify the dependencies among fitness data using the Fitbit dataset.
- Step 2: we will use the second approach to identify the dependencies among other personal data taken from the literature.

These two steps allow to build the BN that will be used in the LooseIT! app example to test the approach on a specific case of inference prediction upon a TP that requires consent for accessing S, A and F user data, as defined in Definition 1, Sec. 4.1.4.

5.2.2 The ground truth dataset

We used a total of 19,817 samples from 49 users. We obtained the samples from the Public data set of Fitbit users from the Open Humans Foundation[3] and the crowd-sourced Fitbit dataset generated from the respondents of a distributed survey via Amazon Mechanical Turk [4] by Furberg et al. As of the time of writing, the first and the second dataset store time series data of 14 and 35 Fitbit users respectively.

The datasets consist of time series data regarding the user's number of steps taken, distance traveled, minutes of activity, floors taken, elevation, activity calories, weight, minutes of sleep, and heartrate information. For some users there is no information about some features, i.e., elevation and floors, which are available only for trackers with an altimeter, and heart rate information for trackers that do not have

[3] https://www.openhumans.org/
[4] https://zenodo.org/record/53894#.W-76oehKjIU

the heart rate monitor. The semantics of Fitbit data that has been used is given in Table 5.1. We considered only the daily data series from the tracker device excluding the manually entered activity values. For example, we considered the resource path activities/tracker/steps (data acquired exclusively from the device) and not the path activities/steps (which includes also the manually entered values when an activity is logged by specifying distance and time) [5].

The datasets were preprocessed, noise was removed (e.g., empty sets when users are not using the tracker), and ranges were normalized resulting in a total of 6,229 clean samples. In machine learning approach, required training sample size varies depending on the problem complexity (e.g, no. of classes and features) and classifier complexity. Following the rule of thumb that a sample size must have at least 50 for small sets features (e.g., 5-10 features) [152] (e.g., the same for bayesian approach in [211]), we conclude that our sample size is enough for this study.

5.2.3 Experimental Analysis and Evaluation

Deriving the structure and the parameters of the BN

Using *bnlearn* as explained in Section 5.1.1, the structures of the BNs that best describe the F data set were learned using 8 out of the 12 standard structural algorithms (4 of the algorithms mentioned in the above section can be only used for discrete data).

For their corresponding parameters, the *bnlearn* fitting function was used using the Maximum Likelihood estimation that is based from the dataset (i.e., 49 users) as explained in Section 5.1.1.

Validation of the BN

After the creation of the BNs, each of them was validated for their effectiveness. It was performed through the k-fold cross validation with $k = 10$. The samples are randomly divided in k subsets and for each round of validation, the remaining subsets will act as the validation set. The chosen loss function will then compute the loss estimates for each round. We used the *bnlearn*'s cross validation function and used

[5] https://dev.fitbit.com/docs/activity/#activity-time-series

TABLE 5.1: Description of the personal data from Fitbit dataset

BN node	Fitbit data	Semantics
steps	activities-tracker-steps	Daily steps counted by analyzing the 3-axis accelerometer data
distance	activities-tracker-distance	Daily distance calculated as walking/running steps per walking/running stride length, where the stride is calculated by height, gender and GPS if available
elevation	activities-tracker-elevation	Daily elevation counted by tracking changes in barometric pressure and movement
floors	activities-tracker-floors	Daily floors calculated as 10 foot increments of elevation
minutesActivity	-activities-tracker-minutesVeryActive -activities-tracker-minutesFairlyActive -activities-tracker-minutesLightlyActive	Calculated as the number of active minutes per day. Active minutes are the average minutes of three levels of activity: very, fairly and lightly active
activityCalories	activities-tracker-calories	Calories burned per day when user is active above sedentary level and excluding manually logged activities.
weight	body-weight and weight-log	Daily measure of weight extracted from per day weightLog and from body time series
minutesAsleep	sleep-minutesAsleep	Minutes in a day in which the asleep pattern is recognized
heartrate	restingHeartRate	The user's heart beats when still. It is estimated from the heart rate taken when the user is awake and asleep.

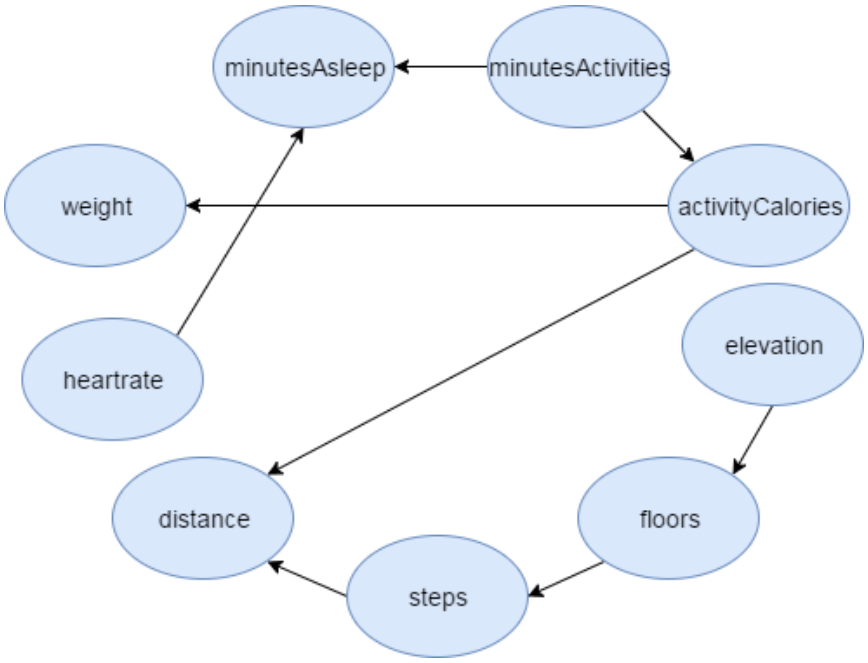


FIGURE 5.1: The chosen BN generated from CHOW-LIU algorithm for the *F* data set.

the Gaussian Log-Likelihood (also known as *negative entropy* or *negentropy*) as the loss function for the validation [161, 199]. Minimizing the negative log-likelihood is the same as maximizing the likelihood. The negative log-likelihood of the validation set is estimated from the BN derived from the training set for each round of validation.

The validation was executed for 100 times to the 8 BNs that were obtained as described in the previous subsection. The expected loss measures are reported in Table 5.2 with their corresponding standard deviation, σ.

TABLE 5.2: BN learning algorithms comparison (negative log-likelihood loss and its corresponding standard deviation, σ, below).

MMPC	SI-HC-PC	HC	TABU	MMHC	RSMAX2	CHOW-LIU	ARACNE
-8.183	-8.185	-10.34	-10.34	-10.3	-10.31	-40.35	-41.2
0.009	0.0089	0.011	0.009	0.0082	0.01	0.379	0.287

Table 5.2 shows that the best candidates for our BN (i.e., those with the highest negative value) are achieved by CHOW-LIU and ARACNE. However, ARACNE does not have a possible orientation of arcs that is fully directed and acyclic which does not make it a BN as stated in Section 5.1.1. Thus, we chose the structure of our BN from CHOW-LIU. The resulting BN is shown in Figure 5.1.

The chosen BN has been further evaluated by running prediction for each node and computing its accuracy. This was done using k-folds of the dataset (as explained above), and was run for 100 times. The Mean Squared Error (MSE) was used to compute the difference between the observed and predicted values for each node in the network. The error percentage and their corresponding standard deviation, σ, has been reported in Table 5.3a. It can be seen that the prediction for the $floors$ node is almost perfect as it is computed from *elevation* by Fitbit. The lowest prediction is obtained by the $minutesASleep$.

TABLE 5.3: Prediction accuracy of each node of the BN

(A) Mean Square Error and its standard deviation, σ, (continuous node case)

weight	distance	mins. Asleep	steps	act. Cal.	floors	heart rate	mins. Act.	elevation
0.0192	0.0052	0.0515	0.0117	0.0002	<0.0001	0.021	0.0072	0.002
<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001

(B) Prediction Accuracy and its standard deviation, σ, (discrete node case)

93.05%	94.16%	61.01%	85.58%	81.89%	99.99%	94.97%	87.22%	99.26%
0.0004	0.0003	0.0006	0.0001	<0.0001	<0.0001	0	0	<0.0001

5.2.4 Discretization of continuous variables

AID-S is a service that computes the risk of inference of data not allowed by user given all the allowed data. In theory, if a user allows AID-S to have all his/her information, AID-S can simply run them on our obtained BN model to compute the risk. However, AID-S is a service that may not be allowed by the user and will not be given the **exact** values of information. Therefore, the BN model must be discretized with two states for every data input, namely, **allow** and **deny**.

Now that we have learned our BN model and evaluated its strength, we aim to extract the maximum probability of inference for each nodes to represent the discretized values. The goal is to present to the user the maximum inference risk of data not shared by the user.

The final Inference Risk Network will only have 2 states for each node (i.e., True = presence of information and False = absence of information) that will represent the maximum probability of true or false inference. The discretization of nodes is also

needed to connect with the other inference studies that can be found in the literature which will be explained in the next subsection.

The discretization of states of the continuous variables was performed to extract the maximum probability of risk (not necessarily the most accurate) that will represent the *True* state. We utilized *bnlearn*'s discretization function that uses interval, quantile, or Hartemik's algorithm[6]. We found that the best case for our variables is to be represented by 4 states which is the case where it gives the maximum prediction accuracy of the nodes as shown in Table 5.3b. These results were computed with 10-folds of the dataset and were run for 100 times. Among all the nodes, it shows that the prediction for the node *minutesAleep* is the weakest and the strongest is again for the $floors$. The results show that this step is coherent with the case of the continuous variables.

After the discretization, we now choose which probability distribution best fits to be the prior distribution for the final Inference Risk Network. We use the theory of maximum entropy, which states that the probability distribution that gives the maximum entropy represents best as the prior, given the uncertainty [80, 147]. We computed the maximum entropy for each probability distribution of nodes, given the different combination of evidences and selected which best represents the network.

Finally, we extracted the maximum/minimum probabilities of the probability distribution to represent the maximum risk probabilities respectively for the *True* and *False* states of the nodes of the final Inference Risk Network. For the *True*, the reasoning is quite straightforward for using the maximum risk. On the other hand, the risk of inference for *False*, given that there is no information, should be less than or (at the worst case) equal to the minimum risk given that there is information. Therefore, since the goal is to represent the maximum risks, we use the worst case for the *False*, which is the minimum risk value.

5.2.5 The complete Bayesian Network

After achieving a BN for the *F* dataset and completely evaluating its strength in the previous subsections, we now connect it with other inference studies found in the

[6] https://cran.r-project.org/web/packages/bnlearn/bnlearn.pdf

literature. The General Inference Network includes the state-of-art inference studies that are related to the personal dataset (S and A), as shown in Figure 5.2.

Upon fusion, it was made sure that the resulting graph was still acyclic (does not loop back to the starting point) and the Markov condition was still satisfied, as stated in Section 5.1.1. This General Inference Network is our complete model with nodes that have *True* and *False* states that represents the maximum risks of inference.

5.2.6 Description of each node

The green, orange, and blue nodes in Figure 5.2 represent the user data sets of S, A and F, respectively, while the black nodes are the inference categories that were found from the literature [31, 217, 142, 200, 140]. Notice that the correlations for S and A have been defined from the literature as well, but they are green and orange colored instead of black since they will be used in the example in the next section. Nodes with two mixed colors represent user data which can belong to both the data sets.

The correlations found in [31] report that the combination of Android smartphone sensor data can infer the user *transportation*, *on-screen finger taps*, and *location*. Another set of inferences found in [217] state that mining the Android public resources gives the possibility to infer several personal data including the user's *location*, *identity*, *disease condition*, and various activities done using the smartphone which includes *investment transaction*.

Health-related inferences are also alarming in relation to the type of data captured by the trackers. The user *heartrate* was shown to have correlation with *blood pressure*, a great predictor for hypertension, and a great risk factor for coronary heart disease (not included in the graph) [140]. Other studies report about the correlation of *strokerisk* with blood pressure and other personal data [200]. *Smoking* gestures are found to be accurately detected (95% accuracy) with wrist-based health trackers [142]. Further inferences, e.g., about *device placement*, *stress*, *conversation*, *emotion*, can be found from the ipShield Inference Database (DB)[7] which were also derived from literature studies.

[7] http://nesl.github.io/ipShield/infdb.html

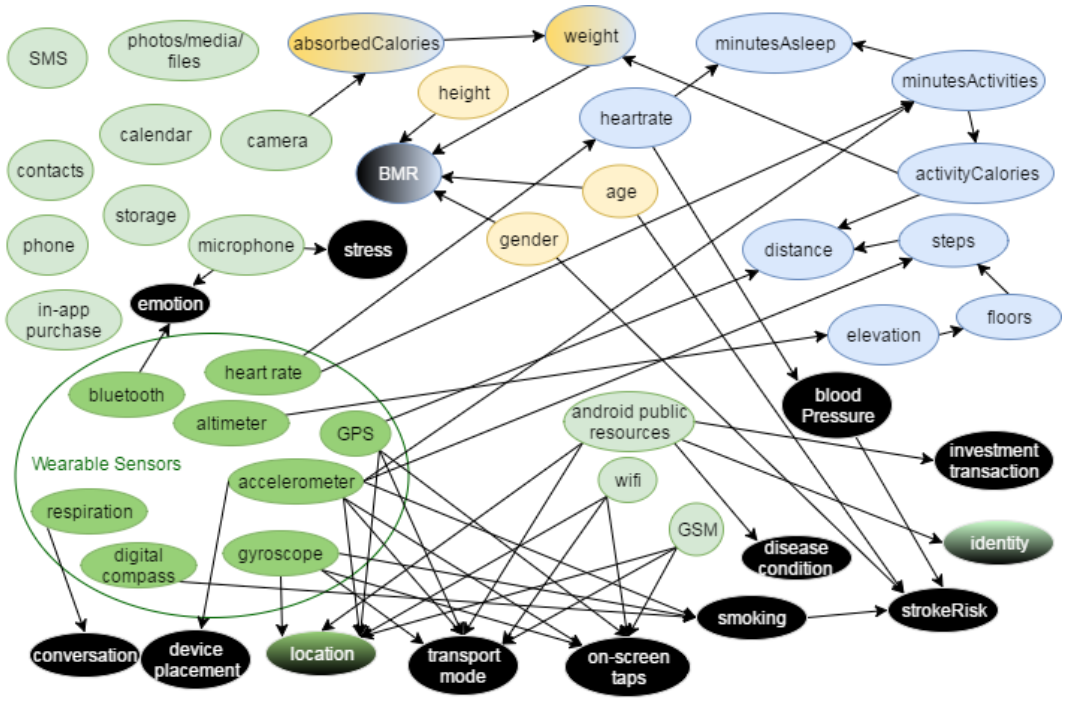


FIGURE 5.2: The general Inference Graph

5.3 Use Case Scenario

This section evaluates the use case formulated in Section 4.1.3 for the LoseIT! TP. First, we show the subset of nodes related to LoseIT!. Then, we provide an evaluation on the strength of this BN subset. Finally, we compute the risks related to the LoseIT! app and the needed recommendation to reduce the related risks.

With respect to the framework, this scenario represents the case where a user decides to install the LoseIT! app. Thus, the AID-S *inference check* module checks for possible inferences, given the set of data that the user has accepted to share with the LoseIT! app.

5.3.1 Subset nodes of the BN involved in the use case

As described in Section 4.1.3, the dataset for the LoseIT! app, TP_i, is composed of $camera \in S_i$, $height$, age, $absorbedCalories$, $gender \in A_i$, $weight$, $activityCalories$, $floors$, $steps$, $distance$, $elevation$, and $minutesActivity \in F_i$. The possible inferences that could be identified from the graph are BMR, $bloodPressure$, $strokeRisk$, and $smoking$. $Smoking$ can be inferred from $strokeRisk$ since the conditional dependency

given by the Bayes' theorem can be computed reversely. The topology is shown in Figure 5.3.

Though *minutesAsleep* is asked in the Lose IT! app, we consider the case when the user denies its permission (i.e., unchecking the 'sleep' box in the permission page) to access her/his *minutesAsleep* information. The aim is to show the possibility for Lose IT! to infer this information given its correlation with the other personal data that the user has allowed to share.

5.3.2 Validation of the subset nodes

As explained in Section 5.2.5, the inference probabilities concerning A and S nodes are based on the correlations from the literature. However, to provide a complete evaluation that includes also these nodes, we used a common method using simulated datasets [60]. To simulate the dataset, we used the correlation metrics for each pair of nodes from the literature and generated a random dataset with the condition that it must satisfy the reported correlation. Thus, we could validate the BN on all the nodes concerning the use case.

The validation is run for 100 times and the result is shown in Table 5.4. The probabilities related to the F dataset did not have a significant change. For the new nodes (from A and S dataset), the prediction for *strokeRisk* and *smoking* have the least accuracy but have overall an average prediction of 72%.

TABLE 5.4: The prediction accuracy and its standard deviation, σ, for each node of the BN subset for the LoseIT! app.

weight	distance	mins. Asleep	steps	act. Cal.	floors	heart rate	mins. Act.	elevation
97.02%	94.16%	61%	85.6%	81.89%	99.99%	94.98%	87.22%	99.26%
0.0002	0.0003	0.001	0.0002	0.00002	0.00004	0	0	<0.0001

camera	age	BMR	stroke Risk	abs. Cal.	blood Press.	height	gender	smoking
77.75%	74.52%	89.39%	50.16%	85.5%	62.11%	74.59%	77.75%	53.92%
0	0	0.001	0.003	0	0	0	0	0

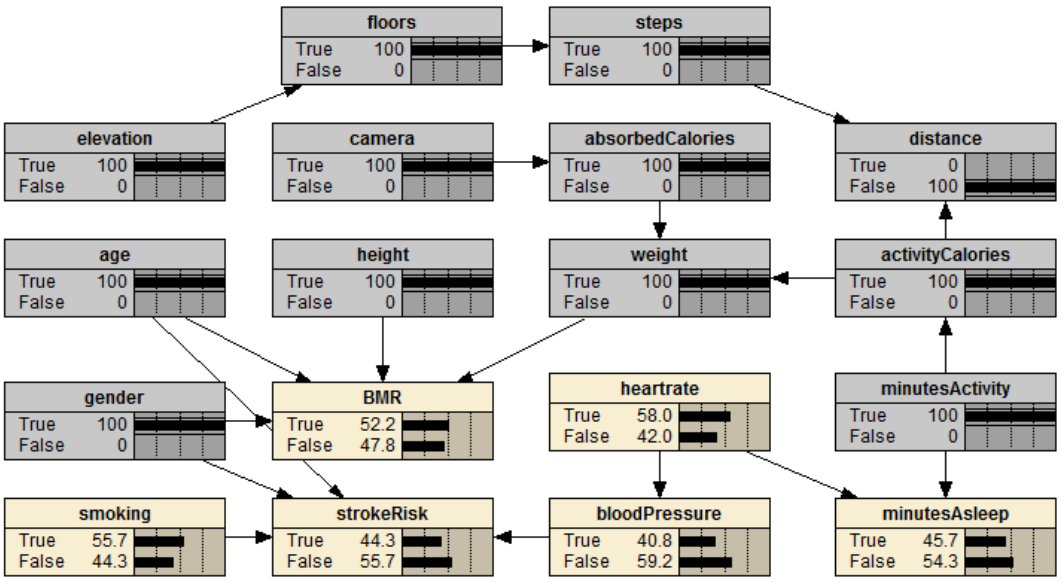


FIGURE 5.3: Inference risks computed for a use case with Lose IT! app.

5.3.3 Risk computation and recommendation

The final Risk Network for the LoseIT! app is shown in Figure 5.3 using the Netica software[8]. The prior and conditional probabilities were estimated from the dataset and the randomly generated dataset using the Maximum Likelihood Estimation as explained in Section 5.2.3.

The personal data asked by LoseIT! app that are allowed by the user were set to evidence *True* in the corresponding nodes (showed as grey nodes in the figure). This propagates the probabilities in the BN, computing the Inference Risk for *minutesAsleep*, *bloodPressure*, *BMR*, *strokeRisk*, *smoking* and *heartRate*.

This becomes a concerning issue since the user does not know that these data items can be inferred by the TP. More importantly, it is alarming in the case of the sleeping information since it was set to private by the user (i.e., unchecked in the permission page). Furthermore, knowing these information may lead to more inference of sensitive information such as the user bed-time activities, health-related information and so on.

The idea of our framework is that the user must be alerted on the possibility that a TP can infer personal data especially those set as private prior to the granting of request and support users in finding suitable alternatives to balance the risk of privacy and the utility of the service.

[8] http://www.norsys.com/netica.html

The implementation of AID-S *Inference check* module described in this study performs the primary task of inference discovery. The second task is to exploit the BN to compute a configuration of privacy permissions that will be able to reduce the inference risk (i.e., *minutesAsleep*) while also taking into account the number of personal data that the user must set to private in order to preserve the utility of the service.

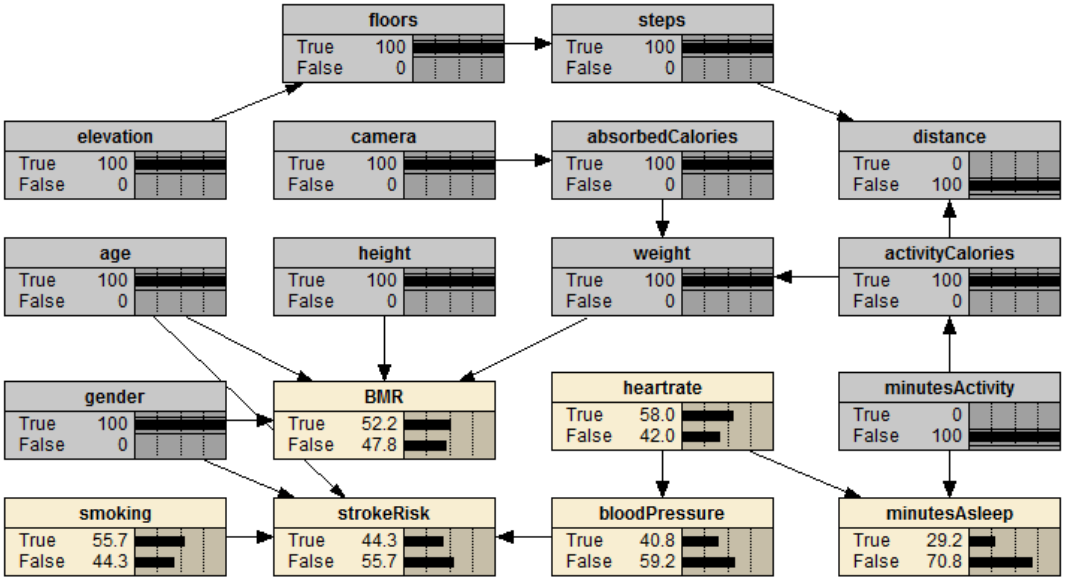


FIGURE 5.4: Recommended Setting for Inference Protection

The objective is to find the minimal configuration of data items that will break the correlation on the private data and, thus, lowers the inference risk.

As shown in Figure 5.4, *minutesActivity* and *heartrate* information are responsible for the possibility to infer *minutesAsleep* information. Therefore, since *heartrate* is not asked by LoseIT!, AID-S can recommend the user to set the *minutesActivity* as private (i.e., unchecking it in the permission page) as an option that optimally reduces the risk. This enables users to take informed decisions and evaluate properly the permission policies according to their privacy preferences.

It is worth noting that in this use case, unchecking the *minutesActivity* is the only possible action of the user given the options available. However, in other conditions, uncheking *heartrate* would also reduce drastically the risks of *minutesAsleep* *bloodPressure*, and *strokeRisk* to 0.10%, 0.44%, 36.1%, respectively.

TABLE 5.5: User Evaluation on AID-S recommendation using χ^2 test.

Fitness Data (A set)	Users w/ 4th party	Users w/out 4th party	All Users
Steps	≈0 (p>0.05)	6.81 (p<0.05)	8.87 (p<0.05)
Distance	≈0 (p>0.05)	11.23 (p<0.05)	13.51(p<0.05)
Elevation	4.29 (p<0.05)	16.04 (p<0.05)	27.33 (p<0.05)
Floors	1.70 (p>0.05)	14.02 (p<0.05)	23.33 (p<0.05)
Calories Activity	≈0 (p>0.05)	6.32 (p<0.05)	5.55 (p<0.05)
Activity Minutes	0.31 (p>0.05)	15.35 (p<0.05)	19.55 (p<0.05)
Heart rate	8.73 (p<0.05)	8.41 (p<0.05)	19.04 (p<0.05)
Food & Water Logs	4.43 (p<0.05)	16.18 (p<0.05)	22.73 (p<0.05)

5.4 User Evaluation on the Recommendation

We also assessed the effectiveness of the AID-S recommendation from the users' feedback. Using the FitPro simulation, as explained in Chapter 4.3.3, we obtained the users' preferences for all permission sets (i.e., S, A, F, G sets). In the subsequent questionnaire, we asked again the respondents for the F Set, but this time with information on risks that are derived from the Inference Graph (i.e., Figure 5.2). The recommendation were simply used as part of the questionnaire and we did not focus on their presentation. We compared the respondents' privacy preferences before (i.e., without the risk information through FitPro) and after (i.e., with the risk information through the questionnaire) using χ^2 test to measure dependency of preference change, where large χ^2 values with small p-values (i.e., p-value < 0.05) indicate dependency on the recommendation. We performed the χ^2 test separately for users who have fourth party/parties accessing their fitness data (e.g., LoseIT! app accessing their fitness data) and for those who do not have. We also performed the test for all the respondents which are shown in Table 5.5.

For the respondents with fourth party/parties accessing their fitness data, results show that the recommendation were significant for **Elevation**, **Food and Water Logs**, and **Heart Rate**. There were significantly more approvals (from previously denied permissions) than denials (from previously allowed permissions) for **Elevation** and **Heart Rate**, while there were equal number of changes (i.e., denials to approvals and vice versa) for **Food and Water Logs**. Given that this group of respondents has

fourth parties, most of the permissions were already approved. The recommendation did not provide statistical significance for the rest of fitness data but it is worth mentioning that after knowing the risks, there were more respondents that allow the permissions which are previously denied than respondents who deny the permissions which are previously allowed. This could mean that, upon knowing the risks, users weigh and approve them since they relate to other fitness data which are needed for the users' fitness goals or the possible information that they can infer are not really private for the users. For the minority respondents, denying the permissions which were previously allowed could mean that the possible data that can be inferred are private for them.

For the respondents without fourth parties, the tests show that they have significant results for all, which mean that there is again dependency on the recommendation. Similarly, there were significantly more approvals than denials after knowing the risks. Given that these respondents do not have fourth parties, they were less likely to give permissions than the respondents with fourth parties. Thus, upon receiving the risk information, more users allow permissions which were previously denied than the respondents with fourth parties. For all the users combined, the results were similar to the respondents without fourth parties, were all recommendation were significant. Interestingly, knowing the risks would make users more likely allow previously denied permissions than deny previously allowed permissions, except for **Food and Water Logs** on the respondents with fourth parties as aforementioned.

Chapter 6

PDM Privacy Preference Model

This chapter discusses the realization of the PDM privacy preference modeling using the Semantic Web approach, which is designed to be GDPR-compliant. The privacy preference model that can be used both by the PDM and TPs in representing the user privacy will be elaborated. The PPIoT Ontology aims to provide general representation of user privacy preferences, which answers **RQ2** defined in Chapter 3.

6.1 Privacy Preference for IoT Ontology

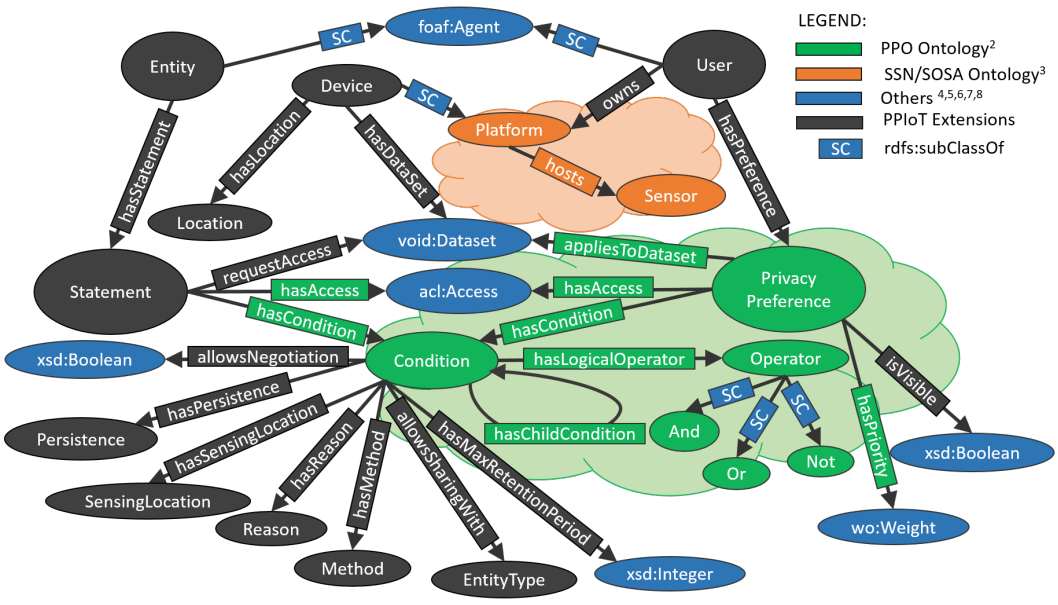


FIGURE 6.1: The proposed Privacy Preference for IoT (PPIoT) Ontology.

The guiding principle for the design of the PPIoT Ontology is that it should be able to represent the conditions of the user's privacy preferences regarding certain personal data (Privacy Preference class) and, on the other hand, the conditions of

the privacy policy statement of a TP entity (Statement class). Thus, the goal is that both the user and the TP, through their respective applications, can set conditions regarding the access to the user's personal data (Datasets class) that are produced by the IoT devices.

The PPIoT Ontology is shown in Figure 6.1. It has been designed to extend existing well-established ontologies with the aim of interoperability. In Figure 6.1, existing ontologies are represented by the green, orange and blue nodes, while the black nodes are the extensions that we propose to cater for the privacy management needs in the context of the IoT.

6.1.1 Main Imported Ontology Classes and Properties

The core ontologies consist of PPO[1] and SOSA/SSN[2]. Only the most relevant classes and properties that are imported from these ontologies are described in this section. Other ontologies that are used include FOAF[3], ACL[4], WO[5], VOID[6] and XSD[7] for defining data types. To avoid ambiguity, in this section a "statement" means an RDF statement while "[privacy] [policy] Statement" is the TP Statement that contains all the details about the data access request. The main imported Classes of the PPIoT are:

- ppo:PrivacyPreference: composed of RDF statements that hold conditions according to the user's personal preferences;
- ppo:Condition: contains the set of properties that denote restrictions to the specific dataset;
- ppo:Operator: The parent logical operator class; the subclasses (i.e., ppo:Or, ppo:And and ppo:Not) allow the creation of more expressive conditions;

[1] ppo:http://vocab.deri.ie/ppo#
[2] ssn:https://www.w3.org/ns/ssn/
[3] foaf:http://xmlns.com/foaf/0.1/
[4] acl:http://www.w3.org/ns/auth/acl#
[5] wo:http://smiy.sourceforge.net/wo/spec/weightingontology.html#Weight
[6] void:http://vocab.deri.ie/void#Dataset
[7] xsd:http://www.w3.org/2001/XMLSchema#

- sosa:Platform: any entity that hosts other entities, actuators, sensors, samplers, and even other platforms. Given the extensiveness of this class, we added a subclass for IoT devices;
- sosa:Sensor: a device, an agent (including humans), or software (simulation) involved in or implementing a procedure. Sensors respond to a Stimulus (e.g., a change in the environment, or input data from the results of prior observations) and generate a result;
- wo:Weight: a value that specifies the priority (rank) of a privacy preference;
- acl:Access: any kind of access mechanism to a resource;
- foaf:Agent: an agent (e.g., person, group, software or physical artifact);
- void:Dataset: a set of RDF statements that describes the sets of data that are collected, generated, maintained, or aggregated by an entity.

The main imported Properties are:

- ppo:hasCondition: the conditions of a privacy preference;
- ppo:hasLogicalOperator: the type of logical operator of the condition/child condition;
- ppo:hasChildCondition: used to create logical nested conditions in combination with the logical operators;
- ppo:hasAccess: the access control privilege which is granted by the user, described using the WAC vocabulary;
- ppo:appliesToDataset: a privacy preference that applies to a Dataset;
- ppo:hasPriority: a value that signifies the rank of a privacy preference;
- sosa:hosts: the relation between the platform and sensor(s).

6.1.2 Extended Classes and Properties

Perera et al. [146] argue that the complexity of the IoT paradigm demands that privacy approaches must offer more than the traditional allow or deny option and instead have room for negotiation between the user and TPs (*Entity* in the ontology). Furthermore, the extended classes must at least be able to represent the list of datasets, the access conditions for both the Statement (for a TP) and Privacy Preference (for the user), the purpose/reason of collection, the persistence of access, the location, the retention period, and the usage method [108, 78, 141, 21]. These principles are also included in the GDPR and FIP. We also included the common data-sharing schemes of TPs where they ask for permission to let other TPs access the user's data as described in Chapter 5 and group them according to Entity type. Below, the new classes and properties that implement the above-mentioned principles are briefly described. The PPIoT [8] prefix are omitted as they are the same for all. The new Classes are:

- User: the owner of the privacy preferences; a subclass of foaf:Agent;
- Entity: any agent that wants to access user information such as a human or a TP application—also a subclass of foaf:Agent but disjoint from User;
- Device: the specific IoT device of the User; a subclass of sosa:Platform;
- Location: the current location of the Device;
- Reason: the purpose of an Entity for accessing the User's data (e.g., health, social, fitness, etc.);
- Persistence: the frequency of data acquisition by the Entity;
- Method: how the data will be processed/utilized;
- SensingLocation: the location of an observation;
- EntityType: the type of Entity (for grouping purposes);
- Statement: the privacy policy Statement declaration of an Entity that consists of conditions regarding the request to access the user's dataset.

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

The new Properties are:

- owns: the relation between the User and her/his Device;
- hasPreference: a privacy preference of the User;
- hasLocation: the Location of the Device;
- hasDataset: the Dataset of the Device;
- hasReason: the Reason of the condition;
- hasMethod: the Method of the condition;
- hasPersistence: the Persistence of the condition;
- hasSensingLocation: the SensingLocation of the condition;
- allowsNegotiation: a boolean data type property that specifies whether the condition (set by the User or Entity) is negotiable;
- hasMaxRetentionPeriod: An integer data type property that specifies the maximum retention period in hours of the data accessed by an Entity;
- hasStatement: the privacy Statement declared by an Entity;
- requestAccess: the Dataset(s) that the Entity request in the Statement;
- allowsSharingWith: which type of Entity is allowed to share the accessed dataset;
- isVisible: a boolean data type property that specifies whether the privacy preference of the User is visible to an Entity.

It is worth noting that both the user and the TP can *set which conditions are negotiable* for the PDM to optimize the negotiation process and recommendation. This can be expressed through the *allowsNegotiation* condition. Otherwise, the setting would be non-negotiable, as is the case in existing ontologies.

Our PPIoT ontology is designed to include classes and properties that address the GDPR requirements for the management of personal data (see in particular the properties *hasReason*, *hasMethod*, *hasPersistence*, *hasSensingLocation*, *allowsNegotiation*,

hasMaxRetentionPeriod, and *allowsSharingWith*). In addition, our interactive PPIoT-based Privacy Preference Model (PPM) meets the GDPR requirements for the communication between the TP and the user, addressing transparency, controllability and accuracy issues.

6.1.3 Ontology Engineering and Validation

PPIoT ontology has been developed following the guidelines and steps stated in Noy and McGuinness [137]. This section discusses the development and validation of the proposed PPIoT ontology.

Domain Modeling and Ontology Definition

Following the guidelines in [137], the first step in developing our ontology was the definition of the domain and scope. While our main goal was clear —i.e., representing privacy preferences from the perspective of both the user and the IoT TP and also taking the GDPR requirements into account—, the definition of the domain model was a non-linear, iterative process. This process started with the collection of relevant domain knowledge and the identification of use cases and competency questions, which were eventually used for the evaluation of our ontology.

In the spirit of the Linked Data paradigm, and following the principle of ontology reuse [171, 137], we also analyzed ontologies that model concepts and relations in our domain of interest, i.e., the IoT domain and the privacy domain. For the IoT domain we consulted the Linked Open Vocabularies for Internet of Things catalog (LOV4IoT[9]), which includes 510 ontology-based research projects in different IoT domains. A recent study [136] showed that, among the IoT ontologies, the W3C Semantic Sensor Network (SSN) ontology is the most commonly re-used ontology in other ontologies and can be considered as a de-facto IoT standard ontology. The SSN (that we selected as our IoT base ontology), also provides alignments to a variety of related ontologies and specifications.

With regard to the privacy domain, we queried the privacy research literature as well as the Linked Open Vocabularies (LOV[10]), which includes stable, high-quality

[9] https://lov4iot.appspot.com/
[10] https://lov.linkeddata.es/dataset/lov/

ontologies. In addition, we consulted experts in privacy legislation. It is worth noting that at the time of designing the PPIoT ontology, while we found 8 vocabularies modeling privacy-related concepts (including the PPO ontology that we selected as our privacy base ontology), we did not find any GDPR-based vocabularies, since drafts of these were published after we developed our ontology.

The PPO ontology was selected since it already modeled user privacy preference for linked data, with a particular focus on social networks. Concepts and relations modeled therein fitted well within privacy modeling requirements in the IoT domain, even though extensions were needed.

The important terms of the domain were then enumerated [137]. Starting from the terms in the PPO ontology, we identified the missing terms for the IoT domain with respect to our goal. Our approach was to extend PPO by defining alignments with other ontologies, and by introducing new terms when none were available from other ontologies. The principles relating to personal data protection are explicitly stated in Article 5 of the GDPR Regulation [51]. These principles also guided the formulation of our *competency questions* in terms of the capability of the ontology to represent, and thus identify, the user privacy preference conditions for different datasets in terms of: lawfulness, fairness and transparency, purpose limitation, data minimization, accuracy, storage limitation, integrity and confidentiality.

Our PPIoT ontology is designed to include classes and properties that address the GDPR requirements for the management of personal data (see in particular the properties *hasReason*, *hasMethod*, *hasPersistence*, *hasSensingLocation*, *allowsNegotiation*, *hasMaxRetentionPeriod*, and *allowsSharingWith*).

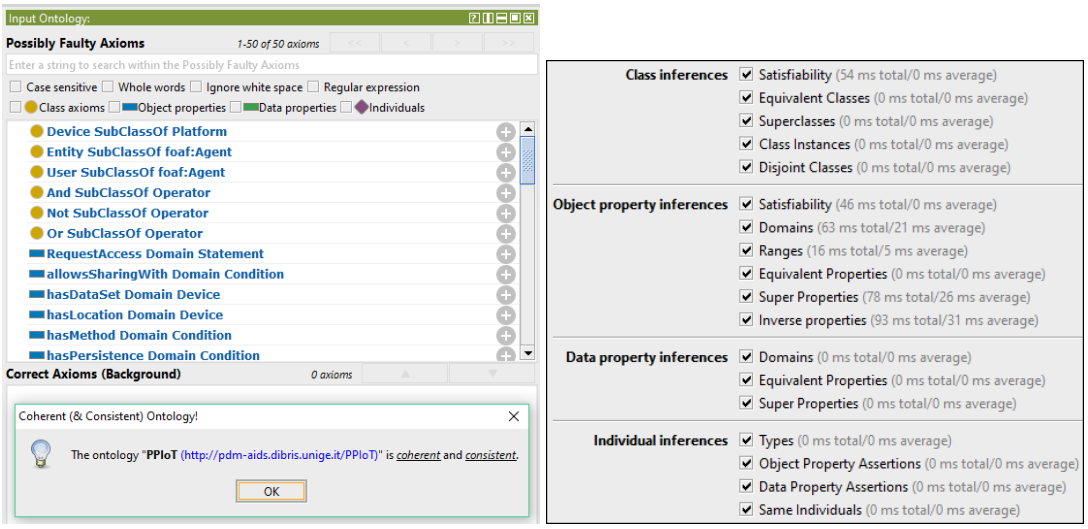
The open-source ontology editor and framework Protégé[11] [129] was used to build the PPIoT ontology.

Ontology Validation

We evaluated the PPIoT ontology by using three methods [137, 75]: (i) a coherence and consistency check, (ii) a task-based and application-based evaluation, and finally (iii) an evaluation using Competency Questions.

[11] https://protege.stanford.edu/

(i) Coherence and consistency check. The studies in [65, 190] state that consistency validation refers to checking whether it is possible to obtain contradictory conclusions from valid input definitions: an ontology is logically consistent when it involves no logical contradiction. The PPIoT ontology was evaluated using different Reasoners in Protege. Reasoners provide consistency checks on the ontology, verifying that the ontology is logically consistent. In our PPIoT ontology, none of the



(A) The result of ontology evaluation.

(B) The inferences tested by the HermiT reasoner in Protégé.

FIGURE 6.2: The PPIoT ontology evaluation in Protégé.

classes and axioms had logical contradictions. Figure 6.2a shows that our proposed ontology is proven to be coherent and consistent using several Protégé reasoners (i.e., FaCT [183], HermiT [163], and Pellet [167]). A total of 50 axioms that are used in PPIoT ontology were tested. We evaluated different reasoners since each reasoner performs differently for each task. In our case, all of them concluded that our ontology did not have inaccuracies. Figure 6.2b shows the inferences and results using the Hermit reasoner, together with its computation time.

(ii) Task-based and application-based validation. According to [75, 23], this type of validation involves evaluating how effective an ontology is in the context of a task or an application. In this light, the "application" may be an actual software program or a use-case scenario [75].

Application-based evaluation has been used, for example, to validate the PPO Ontology [156]. Sacco and Passant validated the ontology by building a a privacy

manager that could implement the creation of privacy preferences for RDF data described using PPO, and that could filter requested data by applying the preferences. Porzel et al. [148] also proposed this method and measured the performance by comparing it to a gold standard.

In our case, we evaluated our ontology by using it in our privacy manager (PDM) to perform the management of users' privacy preferences. The PDM prototype can be found online[12]. It has been tested and has been found to satisfy the requirements to model users' privacy preferences and the TPs' request statements for user data. This prototype will be addressed in Section 6.2.3.

In addition, we indirectly evaluated the PPIoT ontology by integrating it into our mock application. The mock application's understandability, control, simplicity and preferability were evaluated by real users on a 7-point Likert scale. The complete details of this evaluation are explained in Section 6.3.2.

While the aim of task/application-based evaluations is not to assess the generalizability of the ontology, but the performance of the ontology to support some tasks, generalizability could be addressed by applying this type of evaluation to more tasks in different applications.

iii) Evaluation by using Competency Questions. Competency questions can be used to design and then evaluate an ontology [137]. For example, this evaluation technique was used for the validation of the OSHCO ontology [162]: the authors developed competency questions for different use cases with the guidance of domain experts and then queried the ontology and checked the correctness of the retrieved answers. In our case, the identification of use cases and competency questions guided the design process, and they were used to validate the ontology as well. Our aim was to design an ontology that is able to answer questions about the management of user privacy preferences in IoT for different types of personal data and for different IoT domains. TP preferences for data requests made to users were modeled in the same way as user preferences. Thus, the two main *competency questions* that we aim to answer through the PPIoT ontology are the following:

- What are the user's privacy preference conditions for his/her different personal data (datasets, according to PPO terminology)?

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

- What are the TP privacy conditions required by TP in their requests of personal data made to users?

More specific questions are aimed at identifying privacy conditions with respect to GDPR requirements (as explained in Section 6.1.2, from the side of the user and the TP), and to identify the user privacy preferences that are visible, i.e., can be queried by TPs. These questions can be answered by using SPARQL queries, as shown in Section 6.2.3 (Listings 6.3 and 6.4).

Ontology development is necessarily an iterative process and this process of iterative design will likely continue through the entire life-cycle of the ontology [137], in order to capture the domain changes and/or to align the ontology with new or updated ontologies that are modeled for that domain or sub-domains.

6.1.4 PPIoT Ontology Running Examples

This subsection demonstrates how to use the PPIoT Ontology to set conditions for both the user privacy preference and the TP statement.

User Privacy Preference

Listing 6.1 is an example of a privacy preference condition, *myPref*, in Turtle[13] notation. A user may have several conditions for different datasets. In this specific example, we show a privacy preference for a user that applies to their *activity* dataset. Prefixes in Listing 6.1 are defined in Section 6.1.

The user preference example has conditions which state that the data access can happen only be once (persistence of access), the maximum retention period of the data is 24 hours, and the data is required to be encrypted if used for fitness-related reasons. These conditions are combined by the *LogicalOperator* ppo:And and can be negotiated (allowsNegotiation = true) except for the last condition (allowsNegotiation = false). Notice that the last condition was expressed through a child condition with the *LogicalOperator* ppo:And (which is used as in [157]). This shows the significance of child conditions in letting users be more expressive with their privacy preferences. Moreover, *myPref* has "read" and "write" access permission, it has the

[13] https://www.w3.org/TR/turtle/

maximum priority (value=1) that rules out other privacy preferences, it allows sharing such dataset to social networks (with friends only), and it is visible to "enhanced" TPs that utilize the PPIoT Ontology so that they can adjust their parameters to conform to the user's conditions (if they are negotiable for this dataset).

```
@prefix ppiot:<http://pdm-aids.dibris.unige.it/PPIoT#>.
@prefix up:<http://www.userpreferenceExample.com#>.
up:userCond1 a ppo:Condition. up:userChildCond1 a ppo:Condition.

up:myPref a ppo:PrivacyPreference;
  ppo:appliesToDataset ppiot:activity;
  ppo:hasCondition up:userCond1
    [ppo:hasLogicalOperator ppo:And;
    ppiot:hasPersistence ppiot:once;
    ppiot:hasMaxRetentionPeriod 24; #xsd:integer
    ppiot:allowsNegotiation true; #xsd:boolean
    ppo:hasChildCondition up:userChildCond1
      [ppiot:hasReason ppiot:fitness;
        ppiot:hasMethod ppiot:encrypted;
        ppo:hasLogicalOperator ppo:And;
        ppiot:allowsNegotiation false; #xsd:boolean];];
   ppo:hasAccess acl:Read, acl:Write;
   ppo:hasPriority wo:1;
   ppiot:allowsSharingWith ppiot:socialNetworkFriends;
   ppiot:isVisible true. #xsd:boolean
```

LISTING 6.1: User preference condition

Third Party Policy Statement

Listing 6.2 shows an example of a statement of an enhanced TP. It shows that the TP would like to request access to the user's *activity*, *sleep* and *heart rate* datasets. The TP requests access to this data for fitness purposes, and promises to store the data encrypted, for a maximum retention period of 24 hours. These conditions are combined by the *LogicalOperator* ppo:And and are negotiable. However, the TP requires

continuous access to this data (persistence), and it does not allow negotiation on this particular point. Finally, it only requests "read" access to this data.

The automatic negotiation of privacy preferences based on the user's privacy preferences (Listing 6.1) and the TP's policies (Listing 6.2) will be described in Section 6.2.3.

```
@prefix ppiot:<http://pdm-aids.dibris.unige.it/PPIoT#>.
@prefix tpb:<http://www.TPExample.org#>.
tpb:cond1 a ppo:Condition. tpb:childCond1 a ppo:Condition.

tpb:statementB a ppiot:Statement;
  ppiot:RequestAccess ppiot:activity, ppiot:sleep, ppiot:heartRate;
  ppo:hasCondition tpb:cond1
      [ppo:hasLogicalOperator ppo:And;
        ppiot:hasReason ppiot:fitness;
        ppiot:hasMethod ppiot:encrypted;
        ppiot:hasMaxRetentionPeriod 24; #xsd:integer
        ppiot:allowsNegotiation true; #xsd:boolean
        ppo:hasChildCondition tpb:childCond1
          [ppiot:hasPersistence ppiot:continuous;
            ppiot:allowsNegotiation false; #xsd:boolean];];
   ppo:hasAccess acl:Read.
```

LISTING 6.2: TP statement

6.2 PDM Privacy Preference Model

6.2.1 Privacy Preference Model for Interactive Privacy Setting

Now that we have described the PPIoT ontology, we present our approach to support the user and the TPs to manage privacy preferences using our PPIoT-based Privacy Preference Model (PPM), showed in Figure 6.3.

The PPM model is composed of:

- the PPIoT ontology which formally specifies the privacy conditions that have to be taken into account when managing and processing personal data. Basically, it addresses the "data management and processing requirements" described in Section 2.5 and reported in Figure 6.3,

- implementation strategies that address the "communication and transparency" requirements. Basically, they consist in an *interactive approach* that requests users explicit consent to each condition specified in the TP statement based on the PPIoT ontology.

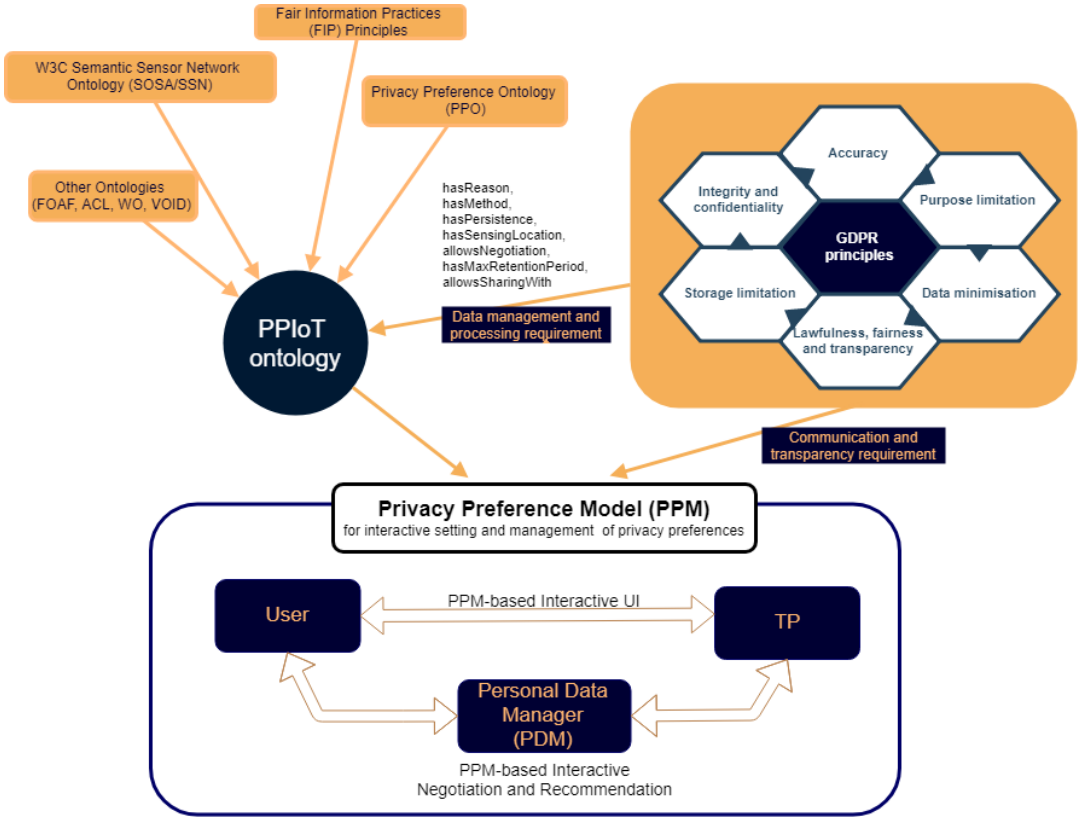


FIGURE 6.3: Adoption of the PPIoT-based PPM in different technological 'regimes'.

.

Figure 6.3 shows that the model can be applied as an interactive user interface or as an interactive negotiation and recommendation process managed by a PDM.

The former approach, "PPM-based Interactive UI" in the figure, can be adopted by TPs for direct interaction with the user. We have developed a mockup of this approach for a fitness application[14]. This mockup improves upon traditional policy

[14] http://pdm-aids.dibris.unige.it/simulation.php

statements, which are usually presented in complex and lengthy "terms and conditions" that users rarely actually read [174, 51]. The PPM-based Interactive UI is based on the PPIoT and is aimed to supplement the traditional privacy policy. The interaction design and layout of the mock-up follow the GDPR requirements of transparency, simplicity, and explicit consent for each request. The TP can store the user's consent data in an RDF store or any other database.

More technologically advanced TPs and user applications can adopt the second approach and use the PPM through a Personal Data Manager (PDM) that conducts an interactive negotiation and recommendation of privacy preferences between users and TPs. The goal, in this case, is to further simplify the management of privacy preferences, relieving the user from the burden of specifying her/his preference conditions for each new device and application, but maintaining the control.

The focus of this thesis is on the interactive settings through the PDM (which will be described in the next section), but it is worth noting that the interactive user interface is based on the same PPIoT-based PPM.

6.2.2 PDM Negotiation and Recommendation of Privacy Preferences

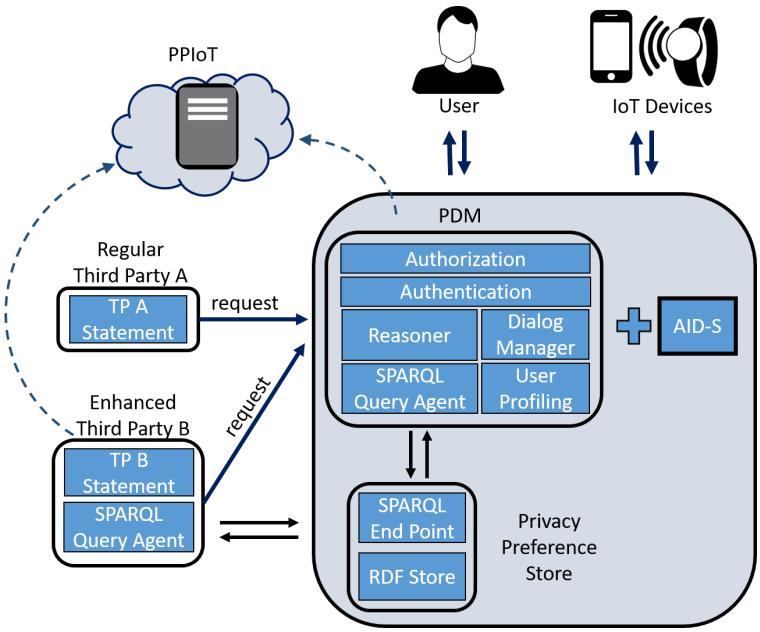


FIGURE 6.4: Overview of the framework implementing the Privacy Preference Model.

We now turn to the use of our PPIoT Ontology in the most technologically advanced scenario, which is the PPM-based interactive negotiation and recommendation through a Personal Data Manager (PDM) inside the PerNANDO Framework.

In this section, we aim to present the PDM's adoption of the SW approach for the management of the user's privacy preferences and the interaction between the user and the TPs. Figure 6.4 shows how the PDM acts as an intermediary between the user, his/her personal IoT devices, and the TPs. The user's privacy preferences, annotated with the concepts of the PPIoT ontology, are stored in an RDF store and made available to the PDM Query Agent through a SPARQL endpoint.

The PDM can be implemented as a client-server application, as a service in the cloud, or even as a semantic mobile app with a mobile endpoint (for an example implementation of this solution see [213]). The current implementation is a client-server application where the client runs as a mobile app while the RDF store and the reasoner are on the server.

The TPs shown in Figure 6.4 include a regular TP (A) and an enhanced TP (B) that has SW capabilities and utilizes the PPIoT Ontology. For TP A, the dialog manager map its policy statement onto a PPIoT-compliant format. Due to the static nature of its policy statement, no negotiation is possible for this TP, and the "negotiation" (i.e., whether to accept the policy or not) will only be between the user and the PDM. TP B, on the other hand, will be able to benefit from the negotiation capabilities embedded in the PPIoT Ontology, effectively giving users the option to specify individual privacy settings rather than wholesale accepting or rejecting the policy.

6.2.3 Privacy Settings Negotiation and Recommendation

In this section, we will show how the PDM uses the PPIoT Ontology and how it performs the negotiation between the user and the TP. The PDM prototype[15] was developed using the Java programming language, the Jena Semantic Web Framework[16] and an Apache Jena Fuseki SPARQL server for RDF storage and querying the user's preference data.

[15] https://github.com/OdnanOriginal/PDM
[16] https://jena.apache.org/index.html

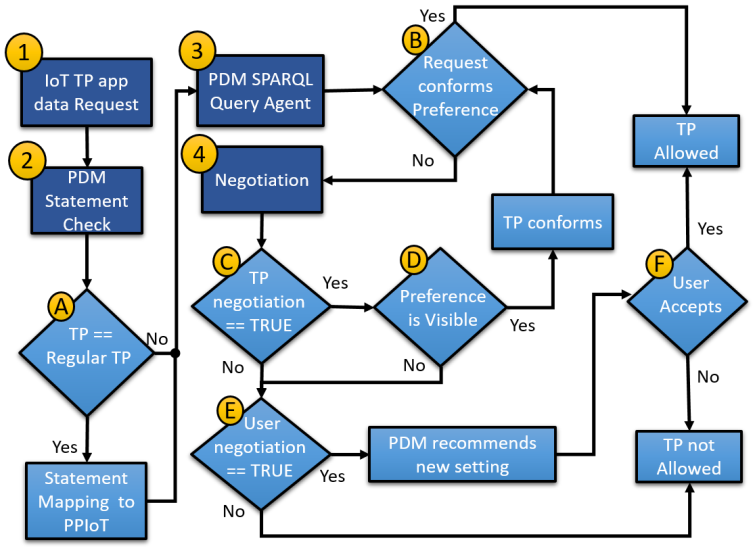


FIGURE 6.5: The simplified interaction workflow between the PDM, the TP and the user.

This use case also provides consistency check on the ontology, verifying that the ontology is logically consistent through the Jena reasoner. None of the classes and axioms had logical contradictions. Figure 6.5 shows a simplified version of the PDM workflow. Its four main steps provide the core phases of negotiation, which will be elaborated below.

TP Application Data Request

In step 1 of Figure 6.5, a TP Statement will be issued during the installation or update of an application. In this instance, the PDM acts as a dialog manager and mediates the interaction between the user and the TP. Our instantiation of the PDM as a mobile application is designed to have the capability to interrupt the installation and check the permissions requested by the TP.

PDM Statement Check

Step 2 is the interpretation of the Statement. In this step, decision block A checks if the requesting TP is a regular TP or an enhanced TP. For a regular TP, the PDM dialog manager can locate its Privacy Statement, which is usually stated in a file (e.g., Androidmanifest.xml[17] for Android apps), and map it onto the PPIoT Ontology. An

[17] https://developer.android.com/guide/topics/manifest/manifest-intro.html

enhanced TP utilizes the PPIoT Ontology to present its Statement, so no mapping is required.

PDM SPARQL Query Agent

Step 3 is the evaluation of the Statement. In this step, the PDM checks if the Statement conforms with the user's privacy preferences. The preferences are queried through the SPARQL Query Agent component. Listing 6.3 is an example of a query from the Query Agent to the privacy preference store that retrieves the list of user's privacy preferences for each dataset. It refers to the example presented in Section 6.1.4.

```
SELECT ?pref ?value
WHERE {?pref <ppo:appliesToDataset> ?value.}
```

LISTING 6.3: A query of the user's privacy preferences for each dataset.

```
SELECT ?cond ?value
WHERE { <:myPref> <ppo:hasCondition> ?tempVariable.
        ?tempVariable ?cond ?value.}
```

LISTING 6.4: A query example of conditions associated to a privacy preference.

By specifying a dataset (e.g., *activity*) for the *?value* variable, the Query Agent can retrieve all user preference conditions associated with this dataset. Subsequently, the PDM can retrieve all associated conditions and their values using the URI name of the privacy preference as input (Listing 6.4). Figure 6.6a shows the result of this PDM query. The query for ppiot:hasAccess, ppo:hasPriority, and ppo:haschildCondition can be done in a similar manner.

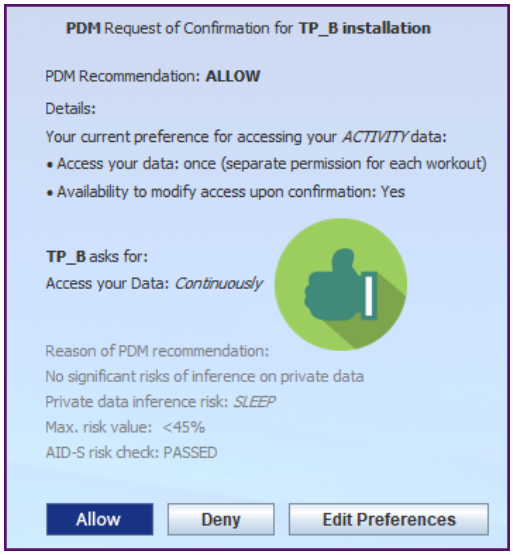
If the TP's request conforms with the user's privacy preferences, it will pass the statement check (decision box B == false) and be granted access. The PDM can now act as an intermediary for the disclosure of the information that is included on the conditions of the user (Privacy Preference) and the TP (Policy Statement) specified in the prior steps.

QUERY RESULTS

	cond	value
1	ppo:hasLogicalOperator	ppo:And
2	ppiot:hasPersistence	ppiot:once
3	ppiot:hasMaxRetentionPeriod	"24"^^xsd:integer
4	ppiot:allowsNegotiation	"true"^^xsd:boolean
5	ppo:hasChildCondition	ppo:userChildCond1

Showing 1 to 5 of 5 entries

(A) The SPARQL query result.



(B) The PDM confirmation request and recommendation.

FIGURE 6.6: The query result of Listing 4 and the PDM recommendation to the user.

Negotiation

If the TP's request does not conform with the user's preferences (decision box B == false), negotiation is needed. There are two cases for this negotiation. If the TP is a regular TP (e.g., TP A), there is no opportunity for negotiation on the TP's side (decision box C == false), and negotiation will only be possible between the user and the PDM if the user has indicated his/her preferences as negotiable (decision box E == true). In this case, the PDM will provide a recommendation to allow the negotiable preference, which the user can accept or deny (decision box F).

Example. Considering the conditions in the example described in Listings 6.1 and 6.2, the PDM finds that the conditions for the TP request to access the activity dataset comply with the user privacy preference regarding the *reason*, *method*, and *maxRetentionPeriod*. However, the *persistence* condition requested by the TP is "continuous" while the user preference is set to "once" for the activity dataset. The PDM finds that the user allows negotiation (decision box E == true), so it recommends giving TP A "continuous" access to the *activity* dataset (see Figure 6.6b). If the user does not allow negotiation (decision box E == false), the TP's request will be denied.

In the case of an enhanced TP (e.g. TP B), both the user and the TP can set negotiation values. The PDM will then first check if the TP allows negotiation on this

aspect (decision box C == true) and if the user's preference for this aspect is set to "visible" (decision box D == true). If so, the enhanced TP can query the user preference using Listings 6.3 and 6.4[18] and modify its request in order to conform with the preferences (decision box B == true). If either the TP does not allow negotiation (decision box C == false) or the user's preference is not set to "visible" (decision box D == false), then negotiation will continue between the user and the PDM alone (decision box E) as described above.

Example. Considering the conditions in the example described in Listings 6.1 and 6.2, upon finding the conflict regarding the *persistence* condition, the PDM will first check if the TP allows negotiation (decision box C) using the algorithm in Figure 6.5. Checking the TP for negotiation first prioritizes the user's preference over the TP request. If the TP allows negotiation (decision box C == true), it will conform to the user's preferences, given that these preferences are visible (decision box D == true). An enhanced TP can query the user preference (Section 6.2.3 shows how to query the preference store using SPARQL queries) if it is set to visible by the user through the *isVisible* property. It can then conform by either removing the request of those data on its Statement that have conditions that conflict with the user's preference or by changing these conditions in accordance to the user's preference. Unfortunately, the enhanced TP B in the example (Listing 6.2) does not allow negotiation (decision box C == false). This could for instance happen if the TP is a fitness tracker, which needs continuous access to activity data to keep track of the user's calories burned. Accordingly, the PDM then checks the user negotiation conditions in Listing 6.1. The PDM finds that the user allows negotiation (decision box E == true). Therefore, the PDM proposes to the user to change the condition for *persistence* to "continuous" (decision box F). Figure 6.6b shows the confirmation request to the user.

In the complete framework, the PDM also calls AID-S for the recommendation to detect whether the request generates any inference risk. In Figure 6.6b, AID-S computation of the inference risk for the datasets asked by the TP is considered to be low (which is shown in gray). If the user thinks the inference risk is too high or anyway if she/he does not agree to grant the permission to the TP, the user can deny

[18]This explains the request arrows from both the PDM and the enhanced TP B to the Privacy Preference Store in Figure 6.4.

the request after all.

6.3 Evaluation

The offline evaluation of the automated negotiation process is done using an actual running application (cf. Section 6.2.3), where an actual use case was simulated and explained step by step. Automation is hard to evaluate in a test environment, though, since it requires simulating all the interaction conditions. Hence, we focus here on evaluating the fully interactive version of PPM, noting that the underlying PPIoT ontology is the same. In this section, we describe the results of our online study with users evaluating the impact of the PPM.

As explained in Section 6.2.1, the goal of the PPM is to provide an interactive approach that enables TPs to use our PPIoT ontology vocabularies to improve control and transparency in the presentation of privacy policy statements. Section 6.2.3 demonstrated how the Personal Data Manager can use the PPM for negotiating and interactively setting the user's privacy preferences. Also, it shows the value and effectiveness of using the SWT by describing TP policy statements and users' preferences with the PPIoT ontology. As Figure 6.3 describes, the alternative means to use the PPM is through an interactive user interface. In this section, we describe the results of a preliminary user evaluation of this user interface. Note that the use of the PPM through the PDM involves an interactive user interface as well, as shown in the Interaction Model in Figure 6.3 —the main difference is that the PDM automates some of the interaction between the user and the TPs.

6.3.1 Interactive User Interface for the Privacy Preference Model

For the evaluation of the PPM-based Interactive User Interface (UI), we used the mockup of a fitness application called "FitPro" which was introduced in Chapter 4.3.3. FitPro presents an interactive PPIoT-based privacy policy to its users.

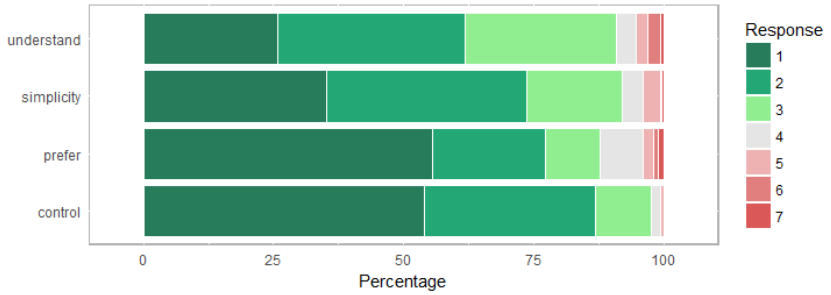


FIGURE 6.7: 7-point scale evaluation on the PPIoT-based PPM.

6.3.2 Sample and methodology

For the evaluation, we recruited a total of 310 Fitbit fitness tracker users via the Amazon Mechanical Turk[19] crowd-sourcing platform linked to our test environment[20]. We restricted participation to fitness tracker users to be able to compare the current privacy settings of their existing fitness app against their settings in our FitPro mockup. Moreover, we restricted participation to Fitbit users to reduce the app-based variability in privacy settings among our participants (cf. each fitness tracker requests slightly different permissions and personal information from its users). Note, though, that our FitPro contains permission requests from a variety of fitness trackers, and hence our results generalize beyond Fitbit to fitness trackers in general.

We removed data from 15 participants whose completion times and answers to the control questions clearly indicated a lack of attention to the study, resulting in a final dataset of 295 responses. The participants are composed of 34.2% males (101 participants) and 65.8% females (194 participants), had a mean age of 35, and were generally highly educated (62% had at least a bachelor's degree).

Participants were asked to use the *FitPro* user interface as if they were installing a new application on their device. The participants were subsequently asked to respond to a questionnaire asking for their feedback about this installation experience, their current Fitbit settings, and their privacy preferences.

[19] https://www.mturk.com/
[20] http://pdm-aids.dibris.unige.it/index.php

6.3.3 Subjective Evaluation

The participants were asked (using 7-point scale items) about the *understandability* of the presented privacy policy, the amount of *control* they thought the interface gave them, how *easy or difficult* is was to set their privacy preferences, and whether they *preferred* this PPM-based interactive settings interface over the traditional privacy policy statements they experienced when installing their apps. Participants' feedback is presented in Figure 6.7, and the following are the averages for each item (lower=better) combined for all respondents:

- Understandability = 2.31 (1=Definitely Understand, 7=Definitely do not understand)
- Control = 1.62 (1=Definitely gave control, 7=Definitely did not give control)
- Simplicity = 2.04 (1=Very easy, 7=Very difficult to use)
- Preferability = 1.86 (1=Definitely Prefer, 7=Definitely do not prefer over the traditional privacy preference model)

The results of our evaluation show that the interactive PPM-based interface helps participants understand the TP's privacy policy, making clear the options and presenting them as a structured, interactive dialogue. Participants tend to prefer this PPM-based interactive privacy policy over a traditional privacy statement, which is unsurprising given that so few users actually read such statements [51, 174]. Arguably, the interactive PPM format is easier for users to engage with, comprehend, and retain. This is reflected in the fact that 85% of survey respondents said they understood the privacy policy. This is a high number in light of the fact that many commercial privacy policies are notoriously hard to understand [81].

Overwhelmingly, the interactive PPM-based interface gives participants more control than a traditional presentation of TP's privacy policy and this is a consequence of the GDPR principles and the interactive presentation. Rather than only being allowed to either accept or reject a policy in its entirety, users can make allow/reject decisions regarding specific aspects of the policy and express conditions. Moreover, despite the complexity that comes with granular control, 90% of the participants find the interface easy to use. Admittedly, users would likely consider it a

burden to make a large number of privacy decisions for a multitude of applications, and/or to frequently revisit these decisions as their privacy preferences evolve. This is where PDM can offer relief in the form of privacy recommendations.

Overall, then, about 80% of the participants prefer the PPM-based interface over traditional privacy policies, with over 50% of participants "definitely" preferring it. Respondents are quite unanimous in their feedback, as standard deviations for these items are low (Understandability: 1.17, Simplicity: 1.25, Preferability: 0.8, Control: 1.07). We also find no significant differences in these evaluations (p-values > 0.05) in terms of gender, age or mobile OS. This shows that GDPR-based PPM does not only conforms to EU requirements for privacy but also results in a more appreciated approach to get consent (i.e., permissions on requested data).

6.3.4 The PPM's Effectiveness on the Elicitation of Privacy Preferences

The above subjective evaluation suggests that our FitPro PPM-based interactive settings are an improvement over the privacy-setting experience of existing fitness apps. The rates at which various permissions were allowed by users in FitPro are displayed in Figures 6.8 and 6.9. Permissions are grouped into the four sets requested in the FitPro simulated installation: In-app requests, Smart phone permissions, Fitness data (Figure 6.8) and GDPR permissions (Figure 6.9).

Our goal in this further evaluation is to compare such FitPro PPM-based permissions against participants' current permissions given to their *existing fitness apps*. In our users' case, the existing fitness apps are: (i) participant's Fitbit app and (ii) the third-party apps on participants' device that request permission to access their Fitness data managed by Fitbit.

It is worth recalling that FitPro simulates the installation of a fitness tracking app like Fitbit. In such installation it requests permissions on personal data (in-app requests), permission to access smartphone data (smartphone permissions), permissions to access, process and store such data (GDPR permissions). Moreover, it includes requests of permissions for sharing FitPro fitness data (Fitness data permissions) with other apps—which simulates third-party apps requesting Fitbit fitness data.

Thus, in principle, we can compare the settings that the participants set in the FitPro PPM-based system against those in their existing fitness apps. However, this comparison is not always possible or meaningful. It is not possible when the participants do not have any current settings (i.e., no third-party app accessing their Fitness app). It is not meaningful when Fitbit settings are mandatory (i.e., mandatory app request permissions and mandatory allow-all blocks of permissions). In both cases, instead of using the fitness apps settings, we asked users about their preferences separately (questionnaire-reported preferences). For each set of permissions, except for GDPR set (this data was not available on participants' devices or even in their experience since it is a novel contribution), we compared the PPM-based settings with the available settings as follows:

- For the *in-app set* (A set), we compare participants' PPM-based settings against their self-reported[21] preferences to adhere to in-app data requests.
- For the *smartphone set* (S set), we compare participants' PPM-based settings against their current settings, i.e., the actual permissions they have given to their Fitbit app (note that the requests differ slightly between Android and iOS).
- For the *fitness data set* (F set), we have two situations: for users who have a third-party application accessing their current Fitbit data, we compare their PPM-based settings against the current settings for one of these third-party applications. For users who do **not** have any third-party applications, we compare their PPM-based settings against their self-reported preferences.

We used chi-square tests to test the association between the settings mentioned above. Tables 6.1, 6.2, and 6.3 show the results of the chi-square tests, that will be explained in the following sections. Large χ^2 values with small p-values (i.e., p-value < 0.05) indicate strong association. Overall, we found that participants' preferences expressed through the questionnaire are strongly associated with their PPM-based settings on FitPro, and the same happens for settings given when the user can freely

[21] We ask for these preferences in our questionnaire because this information is mandatory in all fitness apps, hence, participants' actual disclosure does not necessarily reflect their true preferences.

TABLE 6.1: Chi-square tests of association between PPM settings and participants' preference on in-app data requests.

In-App Request (A set)	Preferences vs PPM settings
First Name	10.6 ($p < 0.05$)
Last Name	11.0 ($p < 0.05$)
Birth date	6.7 ($p < 0.05$)
Gender	1.3 ($p > 0.05$)
Height	0.2 ($p > 0.05$)
Weight	1.4 ($p > 0.05$)

allow or deny each permission, as in Fitbit phone permissions. Conversely, PPM-based settings are not significantly associated to the current settings of all the other third-party apps on the user's device.

The overall results, that will be described later in detail, seem to suggest that the more transparent and controllable PPM-based interactive setting, besides being more appreciated by participants, would also have an impact on the effectiveness in expressing their privacy preferences, even though further investigations are required to generalize these findings.

Below we discuss the results of these association tests for the sets requested in the FitPro simulated installation.

In-App Request Permissions

Fitness apps regularly ask users for their personal data such as name, surname, age, height and weight during sign-up. In most apps, this is compulsory information. In our study, however, we asked the participants if they would allow or deny such permissions if they were optional instead of required. From here, we compared participants' preferences with their PPM-based privacy settings. As depicted in Table 6.1, participants' preferences and PPM settings are strongly associated for first name, last name and birth date. Interestingly, these are also the most sensitive data in this group (see Figure 6.8). For the other items, there is no significant association between participants' preferences and their FitPro PPM-based settings.

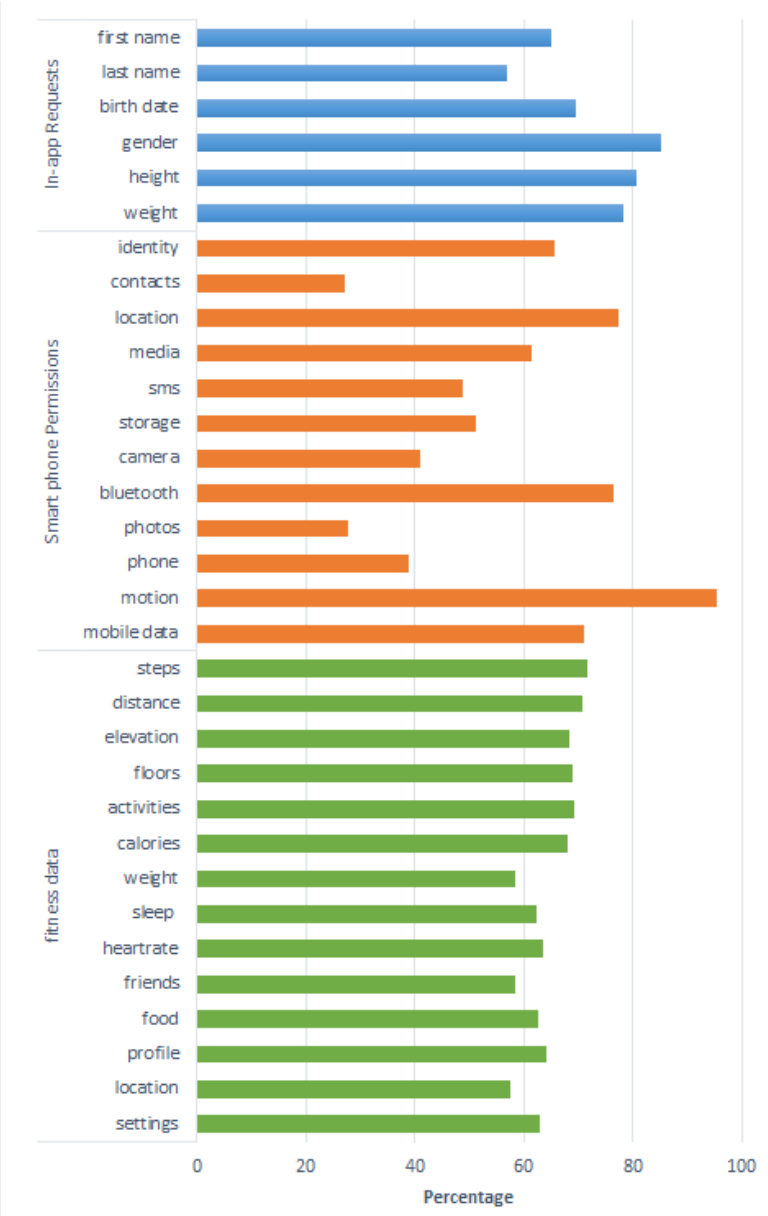


FIGURE 6.8: FitPro permissions allowed by the participants.

Smartphone Permissions

Participants' average acceptance rates for smartphone permissions varies widely; participants were least likely to give FitPro access to their contacts and photos, but most likely to give the app access to their Bluetooth, location, and motion (see Figure 6.8). This is not unexpected for a fitness tracker app.

In comparing against participants' existing settings, we note that Fitbit asks a different set of phone permissions depending on the user's mobile Operating System (OS). Among our respondents there are 162 iOS users, 103 Android 6.0+ users,

TABLE 6.2: Chi-square tests of association between PPM settings and participants' preference on smartphone permissions.

Phone Permissions (S set)	Current settings vs PPM settings	
	Android (6+)	iOS
Phone	8.2 ($p < 0.05$)	-
Storage	9.0 ($p < 0.05$)	-
SMS	17.9 ($p < 0.05$)	-
Contacts	14.6 ($p < 0.05$)	43.2 ($p < 0.05$)
Location	13.5 ($p < 0.05$)	33.8 ($p < 0.05$)
Camera	22.7 ($p < 0.05$)	17.6 ($p < 0.05$)
Bluetooth	-	26.0 ($p < 0.05$)
Photos	-	17.6 ($p < 0.05$)
Media & Music	-	33.6 ($p < 0.05$)
Motion & Fitness	-	10.8 ($p < 0.05$)
Mobile Data	-	37.2 ($p < 0.05$)

17 Android users with an older OS (which does not allow them to control each permission separately), and 13 Windows users. In our evaluation we only consider the two larger groups of iOS and Android 6.0+ users.

Unlike the in-app requests, Android 6.0+ and iOS phone permissions are not mandatory, meaning users can allow or deny each permission separately. We asked participants to tell us their current permission settings for the Fitbit app. Given that participants can freely allow or deny each permission, we assume that their settings are aligned with their preferences. We compare these preferences with the PPM-based settings and the result shows that they have significant statistical relationship for all the permissions for both the Android 6.0+ and iOS, as shown in Table 6.2.

Fitness Data

The fitness data produced by the user's fitness app can be accessed by other TPs to provide more services and features. There are, in fact, many of these external apps that use Fitbit's data. We ask participants to list the current permission settings of the external app they use most. Only 179 participants reported that they had an external app that accesses their fitness data, so for those who do not have an external app, we asked them what their preferences would be for sharing their fitness data with such an app. We report the results for these two groups of participants separately.

Note that, in FitPro, exercise data are broken down into smaller granularity (i.e., steps, distance, floors, elevation, activity minutes, calories burned), giving users more options to control their privacy. However, we generalized these items into a single permission (i.e, Exercise) to be comparable with the settings on the external app that accesses their Fitbit's data.

For participants who have external apps, we compared the current settings of their most used third-party app with the PPM settings as shown in the left column of Table 6.3. It shows that their current permission settings show no association with their PPM settings (i.e., no statistical significance, all $p > 0.05$). It is possible, though, that their current settings do not reflect their real preferences. Indeed, the mismatch between users' privacy preferences and their settings is a phenomenon known as the "privacy paradox", which is well-established in previous research [135, 44, 186, 170, 207].

For participants who do not have third parties, we compared their preferences with their PPM settings, as shown in the right column of Table 6.3. It shows that their preferences are all significantly associated with their PPM settings. This means that participants' preferences on third-party sharing are captured by the PPM for all fitness data items ($p < 0.05$).

Another important thing shown in this figure is that the fitness data show very little variability (see Figure 6.8). This is likely because *what* fitness data shared is less important to the user than *who* the information is shared with (which is part of the GDPR permissions, as discussed below). This result is in accordance with previous studies such as in [108].

6.3.5 GDPR permissions

Unlike the permission sets discussed above, GDPR permissions are a novel contribution of our work, and hence we do not have participants' existing permissions to compare with. Hence, we simply report the results of the GDPR permission settings from our study in Figure 6.9.

For the frequency of access, we let participants choose between granting FitPro continuous access, separate access for each workout (semi-continuous), or only grant access when using the app. Most participants only want to give access when using

TABLE 6.3: The table of chi-squared test of association between PPM settings and participants' preferences on user fitness data.

Fitness Data (F set)	Users w/ TPs: Current Settings vs PPM settings	Users w/o TPs: Preference vs PPM settings
Exercise	0.1 ($p > 0.05$)	7.8 ($p < 0.05$)
Weight	0.2 ($p > 0.05$)	5.9 ($p < 0.05$)
Sleep	0.0 ($p > 0.05$)	6.9 ($p < 0.05$)
Heartrate	0.3 ($p > 0.05$)	15.0 ($p < 0.05$)
Food & Water	0.3 ($p > 0.05$)	12.5 ($p < 0.05$)
Location	0.7 ($p > 0.05$)	11.5 ($p < 0.05$)
Devices & Settings	0.5 ($p > 0.05$)	26.7 ($p < 0.05$)
Friends	0.6 ($p > 0.05$)	27.4 ($p < 0.05$)
Profile	1.0 ($p > 0.05$)	13.3 ($p < 0.05$)

the app, which is a very useful privacy control since apps usually run on background even when not being used [194]. Those who chose to give the app continuous access may want their fitness tracker to count the calories burned and number of steps taken throughout the day, which is one of the main features of many fitness trackers.

For the retention of data, participants are given the following options: store until no longer used, store until the app is uninstalled, or store indefinitely (as to recover during app re-installation). Only 1% of the participants chose the latter. Most of the participants prefer to retain their data until the app is uninstalled (47%) or would like to store it until no longer used (42%). This even split shows that participants have different preferences regarding retention, which means that this permission is important and must be controllable by the user.

The purposes of data collection are then specified. Among our participants, 85% allow data collection for health purposes, 84% for safety purposes, 54% for social purposes, 62% for convenience purposes, and only 17% for commercial purposes. Having the option to deny data use for commercial purposes could solve many privacy issues that stem from commercial disclosure without the user's informed consent. On the other hand, we acknowledge that this is an integral part of many companies' business model.

Finally, GDPR entity types include social media apps, fitness apps, commercial and government fitness programs, and other apps on the user's phone. Fitness tracker users mostly allow sharing to fitness apps and social apps, but the latter

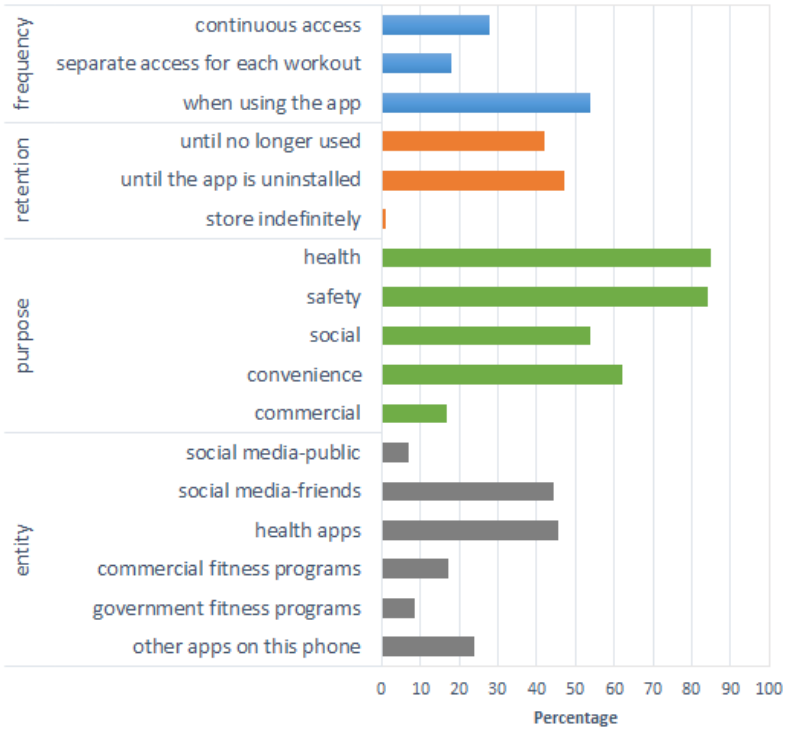


FIGURE 6.9: GDPR permissions allowed by the users.

only if sharing can be restricted to their friends. In fact, participants are least likely to share their data on social media publicly.

In general, our study is among the first to measure users' preferences regarding GDPR-mandated permissions. Our results show a substantial variability in users' preferences regarding these permissions, which is a testament to the importance of these permissions in the fitness domain, and likely beyond as well.

Chapter 7

PDM Privacy Recommendation

This chapter is devoted on privacy recommendation, which addresses **RQ3**,i.e., how can we aid users to set their privacy and provide them with suitable recommendation?

To answer this question, we created different privacy profiles using ML clustering. Then, we used supervised ML to find determiners that can identify which privacy profile best fit for a given user. Finally, we designed different recommendation strategies that interacts with the user to provide the recommendation.

This lengthy task can be broken down into three main parts, each with the following research question:

> **RQ3.1** *Is it possible to identify well-defined privacy profiles that can represent the diversity of users' privacy preferences?*

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

> **RQ3.2** *Are there any privacy profile items or questionnaire items that can be used as a determiner to predict which privacy profile best describes a user?*

To answer this question, we conducted 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 **RQ3.2**.

RQ3.3 *How can we effectively exploit the results to provide recommendation?*

To answer this question we developed a series of recommendation strategies and user interfaces based on the machine learning results. We aim to integrate these recommendation strategies within our PDM framework and aid users by balancing privacy recommendation accuracy and privacy-setting simplification.

This chapter provides in-depth discussion on addressing the formulated research questions above. Chapters 7.1, 7.3, and 7.2.1 discuss the privacy profile clusters, finding predictors to such clusters, and recommendation strategies to address **RQ3.1**, **RQ3.2** and **RQ3.3**, respectively.

7.1 User Profiling Models

In this section, we present our data analysis, demonstrate our method of clustering privacy settings, and generate cluster-based privacy-setting profiles. The analysis in this section is intended to address **RQ3.1**. The data collection for this study has been discussed previously in Chapter 4.3.

7.1.1 Data Analysis

Figure 7.1 shows that there is considerable variability in the average rate at which each permission is allowed or denied in our study. The permissions requested by the application (A set), mainly concerning demographics, have a high disclosure rate, which is in line with the results of other studies (cf. [96]).

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 photos or contacts is granted much less often.

Regarding the purpose, frequency and retention period of data collection (G set), participants seem most 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). On the other hand, 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.

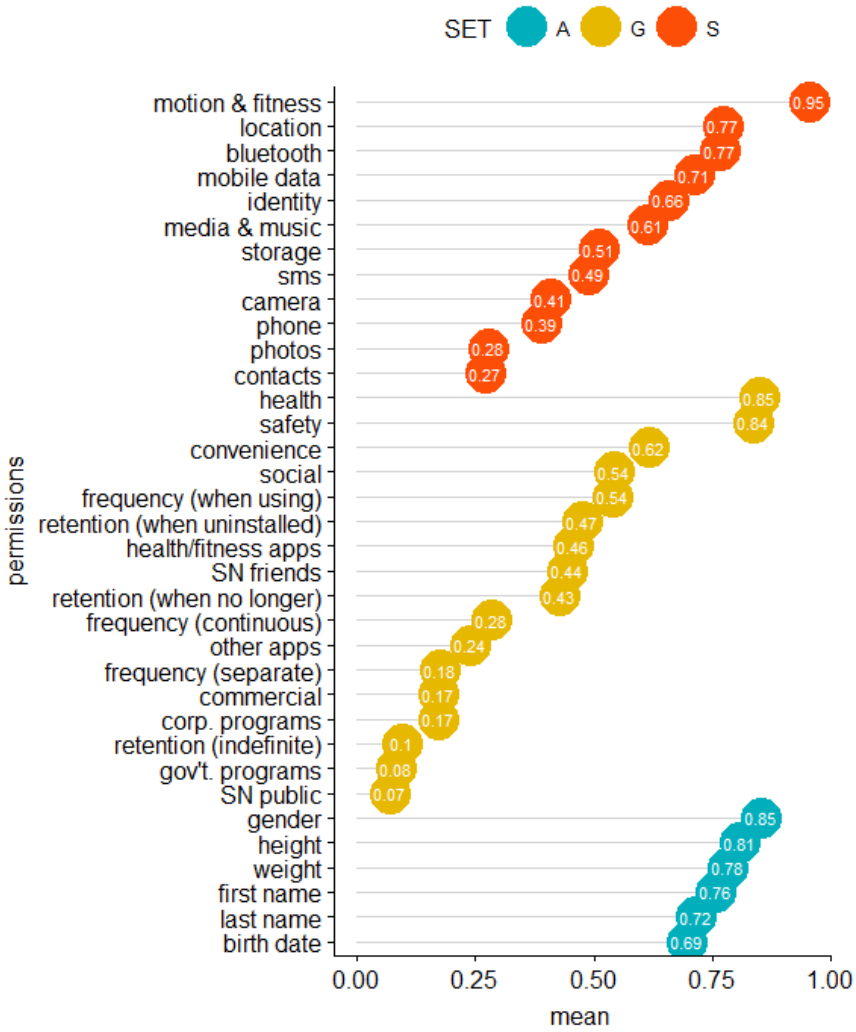


FIGURE 7.1: Average values of each privacy permissions (1-allow, 0-deny).

We do not show the fitness data (F set) in Figure 7.1 because the permissions for these data are requested for multiple entity types of the G set, as discussed in Section 4.2.3. Hence, we present these data in Figure 7.2 instead, showing each permission for each GDPR *EntityType*.

Users are more likely to give permission to 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.

Upon further inspection, we note that participants tend to share either (almost)

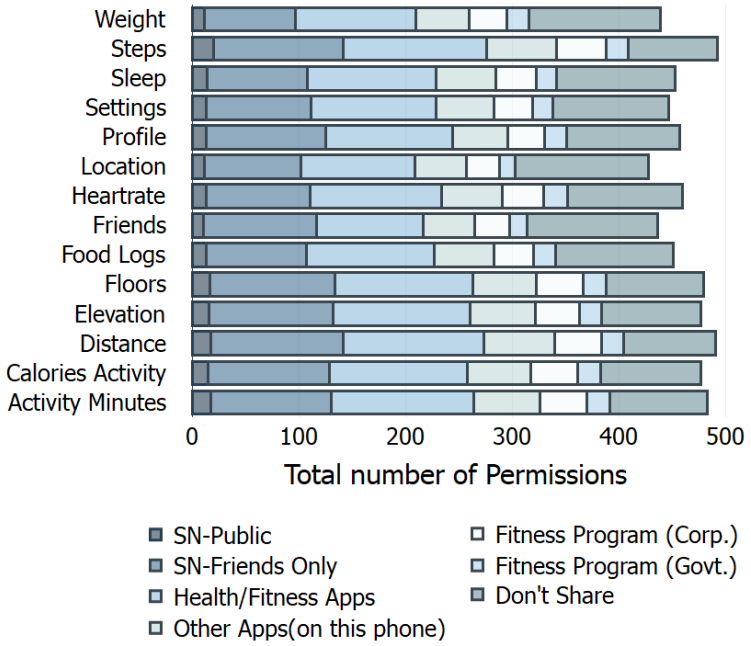


FIGURE 7.2: Fitness data (F set) distribution for each Entity Types (G set).

all or (almost) none of fitness data with an entity. This suggests that Fitness data permissions are more likely to be influenced by the receiver ("who") rather than the specific data item ("what").

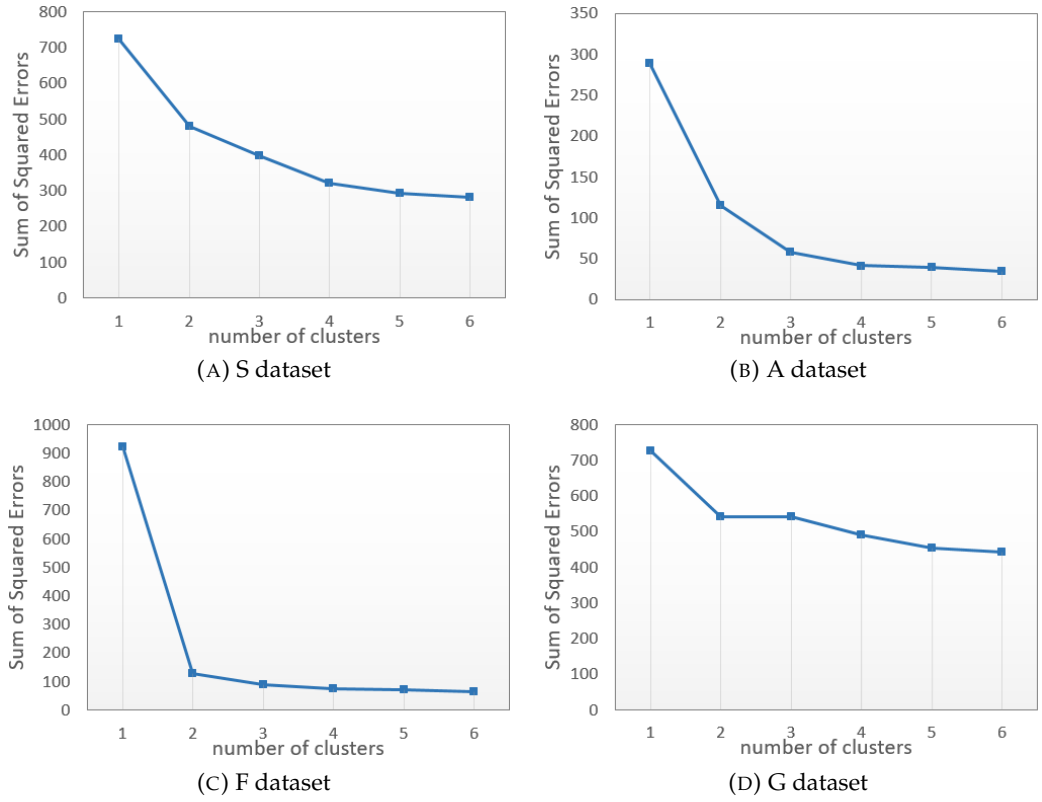
As discussed in Section 6.1, these "who" parameters are instances of the GDPR *EntityType*. Therefore, we expect that clustering F permissions should provide a unanimous deny/share for all items, while clustering G permissions should provide more nuanced clusters of different entity types receiving the data specified in the F set.

7.1.2 Clustering Methods

Our dataset shows considerable variability between participants' privacy preferences—a finding that is broadly reflected in the privacy literature (cf. [93]). 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[1] using the K-modes clustering algorithm [32] with default settings. The K-modes algorithm follows the

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



(A) S dataset (B) A dataset

(C) F dataset (D) G dataset

FIGURE 7.3: Evaluation of different numbers of clusters for each set.

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 full dataset, including the A, F, S, and G subsets. A drawback of this method is that, assume we cluster the users 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 (e.g., the sum of squared error for the viable 4-cluster solution is 1435), for anything but a very large number of clusters.

If we instead generate a separate set of n "subprofiles" for each of the four datasets (A, F, S, 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 Figure 7.3. For example, with only 2 clusters per set, the sum of squared error reduces to 1277 (a 24.3% reduction). An additional benefit is that the profiles for each set can be investigated in more detail.

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.

7.1.3 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 Figure 7.3. Using the elbow method [100], we conclude that 2 is the optimal number of clusters for each dataset[2].

The final cluster centroids of the 2-cluster solution for each dataset are shown in Figure 7.4, together with the results of the 1-cluster solution. We describe the subprofiles of each set in the subsections below.

The *S* Set

- **Minimal** (cluster 0): 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 1): this subprofile allows all permissions in this dataset.

The *A* Set

- **Anonymous** (cluster 0): this subprofile shares only users' gender, height and weight information but not their birth date or first and last name.
- **Unconcerned** (cluster 1): this subprofile shares all data requested in this dataset.

[2]We obtain similar results using other clustering algorithms, such as Hierarchical Clustering.

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

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

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

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

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

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

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

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

FIGURE 7.4: Privacy profiles from the two clustering methods: 1-cluster results (full data) and 2-clusters results (privacy subprofiles) for each dataset

The *F* Set

- **Unconcerned** (cluster 0): this subprofile shares all fitness data with TPs.
- **Strict** (cluster 1): this subprofile does not share any fitness data with TPs.

The *G* Set

- **Socially-active** (cluster 0): 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 1): this subprofile does not allow sharing with any TPs. Sharing is allowed only for health and safety purposes.

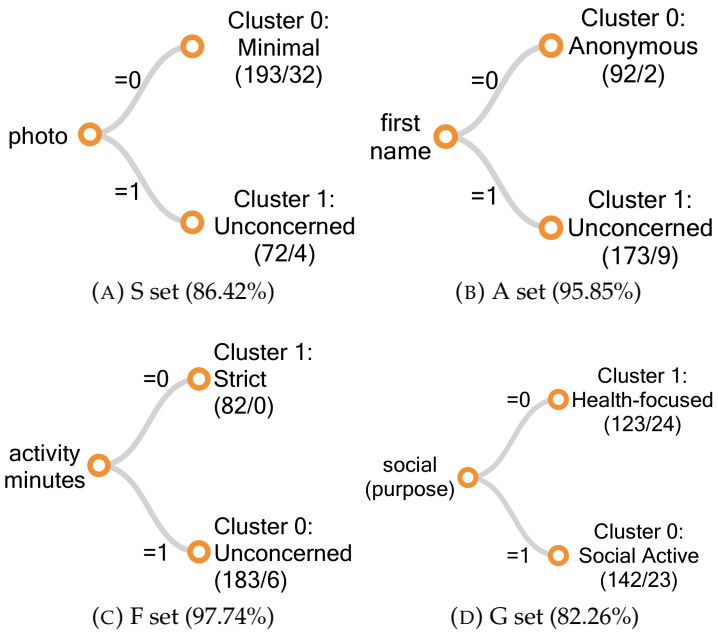
7.2 User Interface Models

7.2.1 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 to. This section aims to answer the research question: **RQ3.2** Are there any privacy profile items or questionnaire items that can be used as a determiner 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 predicition" approach. Additionally, we also explored whether the items from our questionnaire (see Section 4.3.4) could drive the predicition. 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 [143]. Previous work show the effectiveness of this approach to predict privacy settings within each cluster [11]; 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 evaluation of all developed J48 trees was performed using k-fold cross validation.



(A) S set (86.42%) (B) A set (95.85%)

(C) F set (97.74%) (D) G set (82.26%)

FIGURE 7.5: The permission drivers for the privacy subprofiles and their respective prediction accuracies.

Direct Prediction Questions

In our direct prediction approach, the aim is to ask users to answer certain permission items 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 Figure 7.5. It shows for each subset the question 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 [11] 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 Figure 7.5. These single-item solutions already obtained a high accuracy, and their parsimony 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 Photo permission is the best subprofile predictor. This is one of the least-shared permissions (see Figure 7.1), 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, Activity minutes permission is the best predictor. This is one of the most-shared permissions. Around 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.

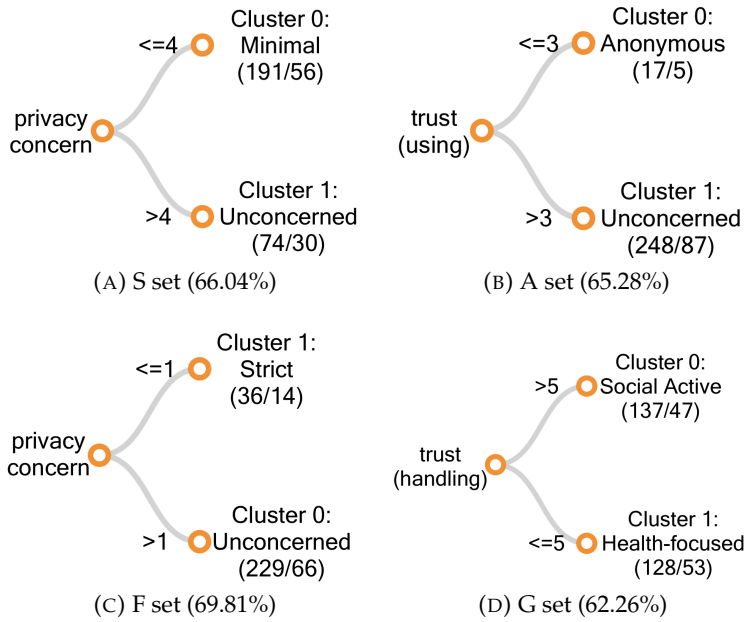
Finally, for the G set, the best predictor is whether the participants allows data collection for Social purposes. If so, participants are correctly classified into the "Socially active" subprofile with 84% accuracy, otherwise they are classified into the "Health-focused" subprofile with 80% accuracy.

Indirect Prediction Questions

A similar procedure was applied to the questionnaire data concerning the following categories of user traits: privacy attitude, social behavior, negotiability, exercise tendencies and user demographics (cf. Table A.1 in Appendix). As will be shown below, the indirect prediction approach has a lower accuracy than the direct approach presented in Section 7.2.1. This is expected since 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. Moreover, the resulting predictors show interesting semantic relationships with the datasets they predict. We discuss these results in more detail below.

Privacy Attitudes

We first attempted to use privacy attitudes as predictors of users' subprofiles. The resulting trees for this indirect prediction are shown in Figure 7.6.



(A) S set (66.04%) (B) A set (65.28%)

(C) F set (69.81%) (D) G set (62.26%)

FIGURE 7.6: The attitude drivers for the privacy subprofiles and their respective prediction accuracies.

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 "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 "Uncocerned".

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.

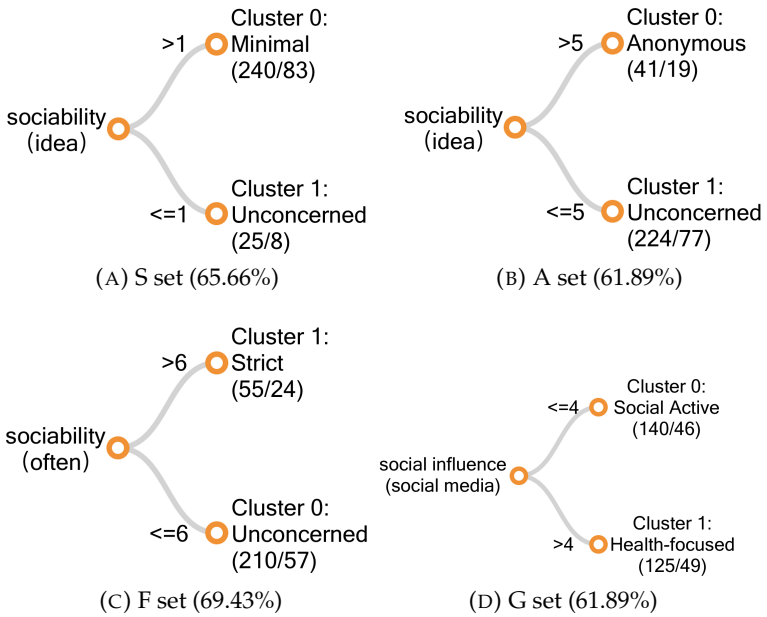
For the G set, those users who agree (6) or extremely agree (7) with the question "I believe the company providing this fitness tracker is trustworthy in handling my

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 result shows that we managed to capture 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 that is one less question to the users compared to the direct questions in Section 7.2.1.

Social Behavior

We also tried to find predictors among the questions about social influence and sociability. The resulting trees for this indirect prediction are shown in Figure 7.7.



(A) S set (65.66%) (B) A set (61.89%)

(C) F set (69.43%) (D) G set (61.89%)

FIGURE 7.7: The social behavior drivers for the privacy subprofiles and their respective prediction accuracies.

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, seem 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.

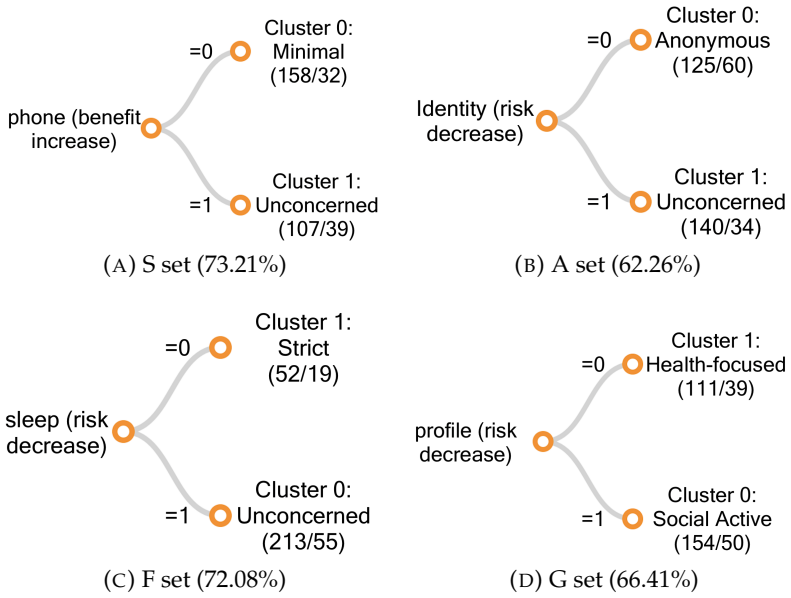
Again, we found interesting semantic relationships between social influence and sociability while exercising and users' privacy-related behaviors: users who are more prone to reap social benefits from exercising are more likely to give the app more widespread permissions. Similar to privacy attitudes,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

Negotiability of Privacy Settings

We also attempted to use the negotiability of users' privacy settings as input for the subprofile prediction. Figure 7.8 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"



(A) S set (73.21%) (B) A set (62.26%)

(C) F set (72.08%) (D) G set (66.41%)

FIGURE 7.8: The user negotiability drivers for the privacy subprofiles and their respective prediction accuracies.

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 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.

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 [96], in this particular study we were not able to find any significant predictors among the group of user demographics.

Tree Evaluation

Figure 7.9 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 $k = 10$.

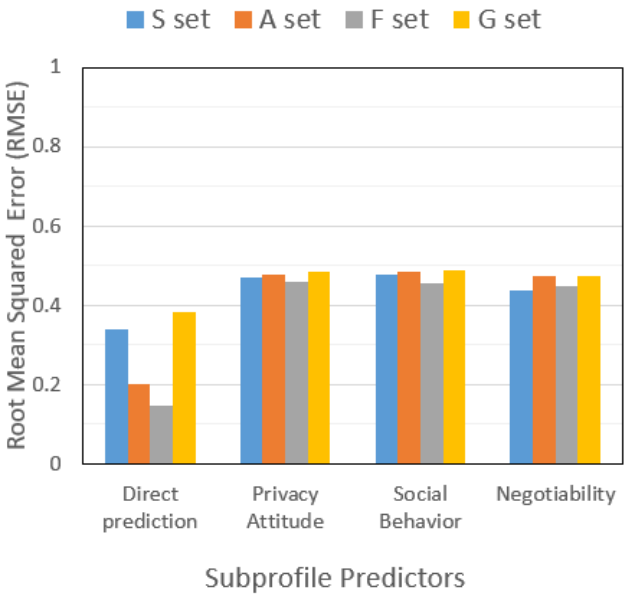


FIGURE 7.9: Evaluation of each J48 tree algorithm on each set.

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).

The prediction accuracies of each tree are reported on the branches in their respective figures (Figure 7.5 to 7.8), and take the form of (# assigned / # incorrect).

7.3 PDM Recommendation Strategies

In this section, we describe different types of guided privacy-setting approaches for IoWT users 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 **RQ3.3**: How can we effectively exploit the results to provide recommendation? We also present a validation of the recommendation results using a hold-out sample of permission settings from 30 additional users.

Privacy-setting Recommendations

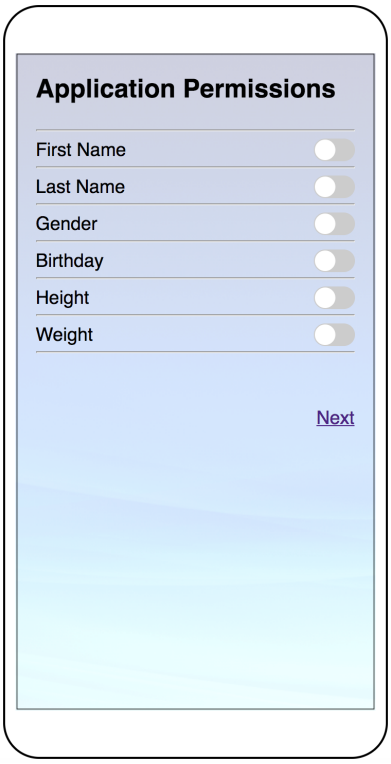
Manual Setting

The baseline privacy settings interface is one where users have to manually set their settings (see Figure 7.10). 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. [122]), so the 100% accuracy may not be achieved through manual settings.

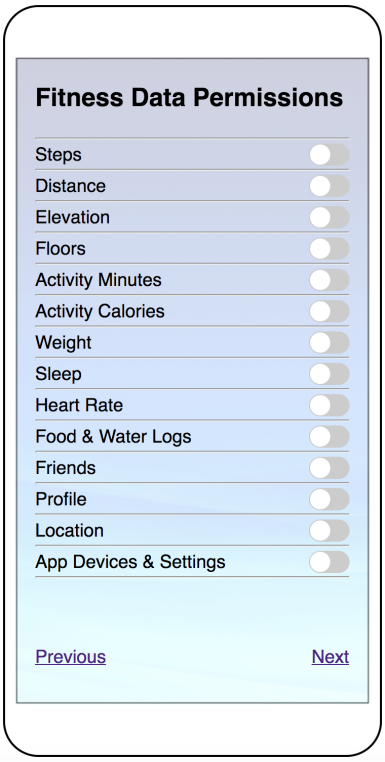
The next strategies exploit the results of the analysis in the previous section to provide *interactive recommendations* that simplify the task of privacy permission setting, with different levels and type of user intervention.

Smart Single Default Setting

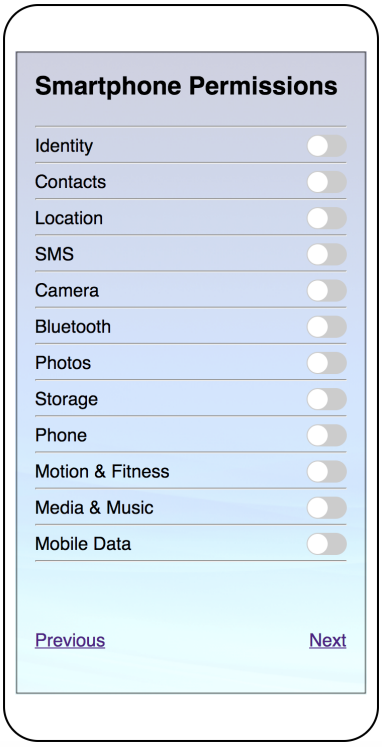
One way to reduce the burden of privacy management is with "smart" Single default setting. Rather than having the user set each permission manually, this solution already selects a default setting for each permission. Users can then review these settings and change only the ones that do not match their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



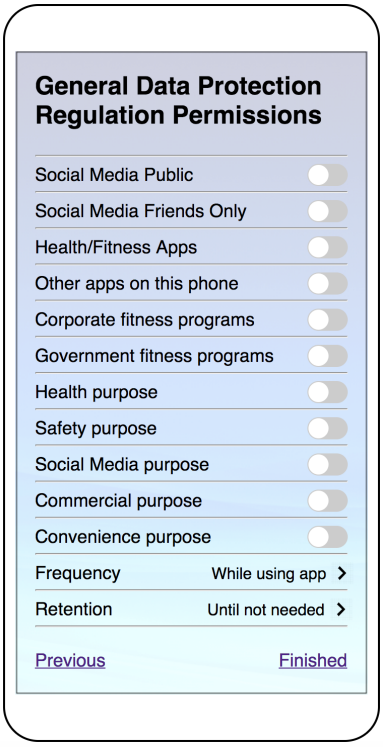
(A) A set



(B) F set



(C) S set



(D) G set

FIGURE 7.10: Manual settings

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 setting by using the cluster centroid of the 1-cluster solution (i.e., the full dataset "single cluster" in Figure 7.4). Figure 7.11 shows the resulting default values for each dataset. If the user is unhappy with these settings, he/she can still make specific changes. Otherwise, he/she can keep them without making any changes.

Pick Subprofiles

The smart single 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. [93]). This bring 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 7.12a shows the subprofile selection interface for the S set. Users can choose either the ‘‘‘Minimal" or "Unconcerned" subprofile, which are shown in Figures 7.12b and 7.12c respectively. Similar interfaces are provided for the F, A, and G sets (not depicted here).

The subprofiles provided by this approach have a higher overall accuracy than the "smart" single default described in Section 7.3, 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").

Direct Prediction

For the direct prediction approach, we devise an interactive 4-question input sequence as shown in Figure 7.13. Each screen asks the user to answer a specific permission question, which guides the subprofile classification processes as outlined in

Section 7.2.1. 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 First name (for the A set recommendation), Activity (for the F set), Photos (for the S set), and whether they allow their data to be used for Social purposes (for the G set). 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.

Indirect Prediction

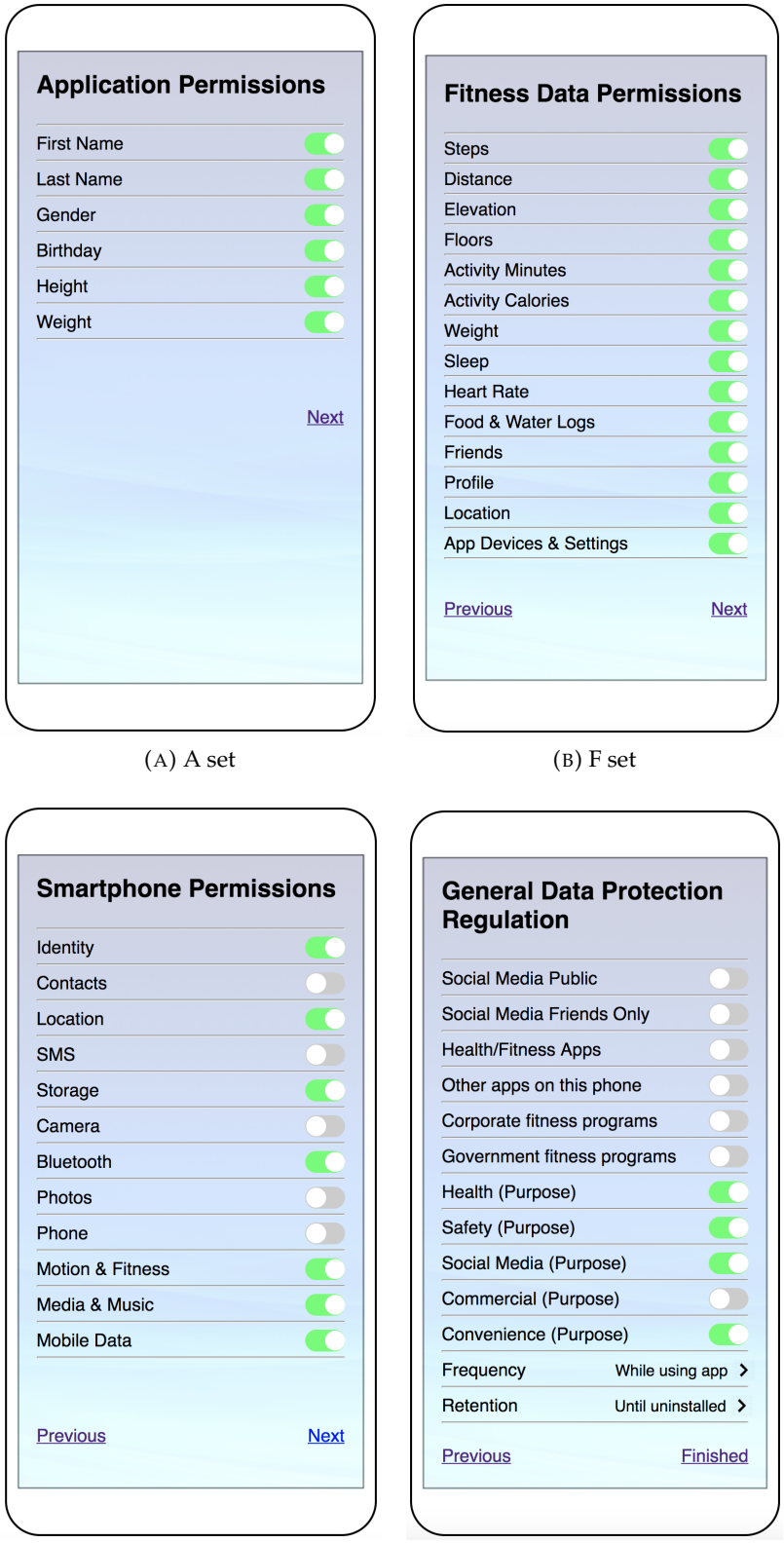
For the indirect prediction approach, we take a similar approach, but the interactive 4-question input sequence is based on the analysis of questionnaire items rather than permission settings.

As shown in Figure 7.14, we selected 4 questions that yield the highest accuracy for each permission set: a negotiability question for Phone permissions for the S set, a negotiability question for the permission to share Sleep data for the F set, A question about sociability for the A set, and a trust question for the G set. Negotiability and attitude have almost the same accuracy for G set, so we chose attitude for diversity.

The benefit of the indirect prediction approach is that the user does not have to answer any permission questions, not even the four needed to give a subprofile recommendation. Instead, the user has to answer four questionnaire items.

Validation

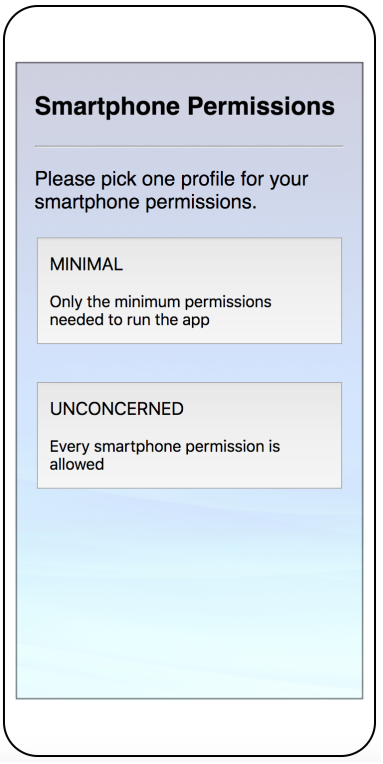
We conducted a validation of these different approaches by running the recommendation strategies on the 30 users in our holdout dataset. The resulting recommended privacy subprofiles are then compared with their actual privacy preference. Figure 7.15 shows the average accuracies of each of the presented approaches.



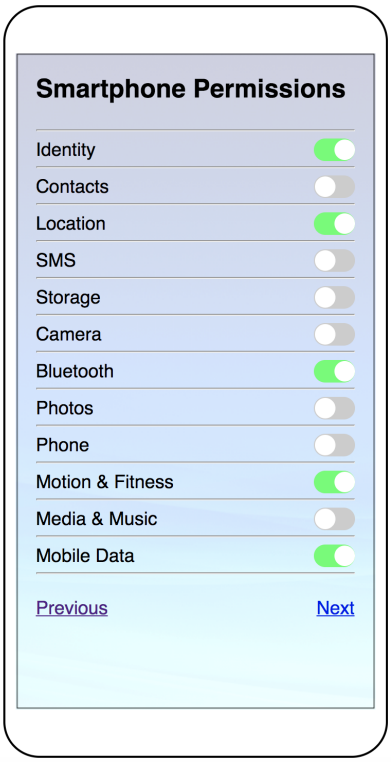
(A) A set (B) F set

(C) S set (D) G set

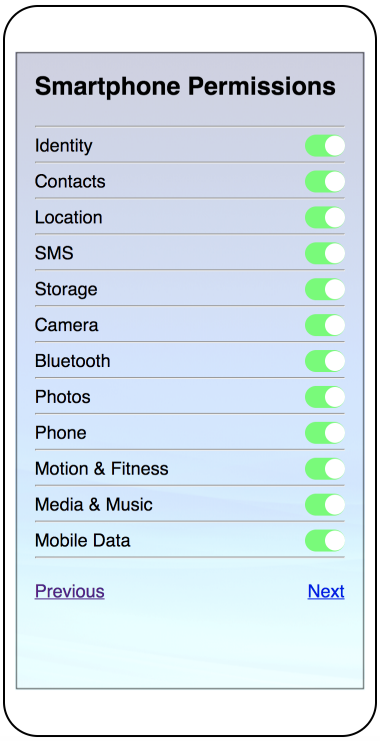
FIGURE 7.11: Smart Single settings.



(A) S set subprofiles

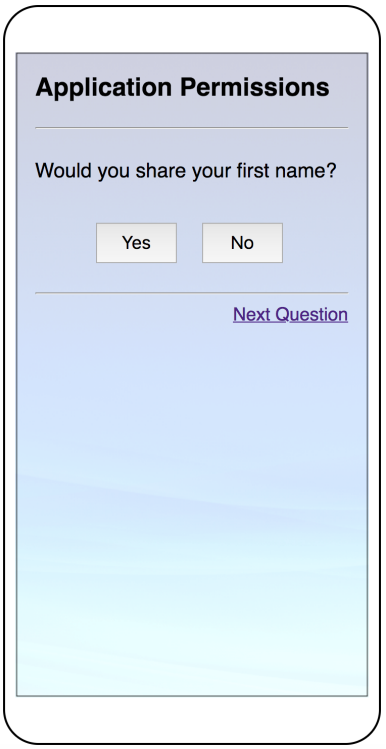


(B) The "Minimal" subprofile

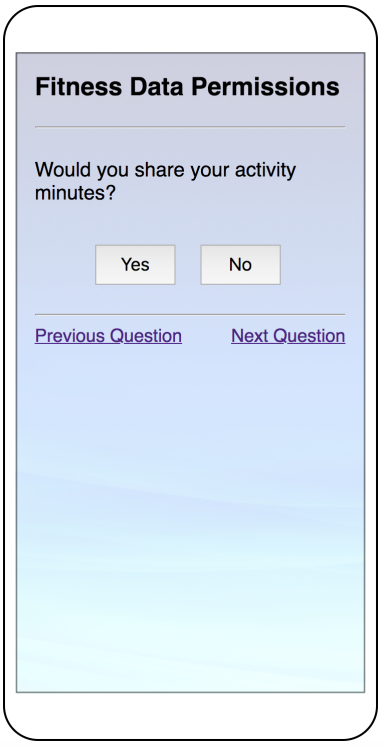


(C) The "Unconcerned" subprofile

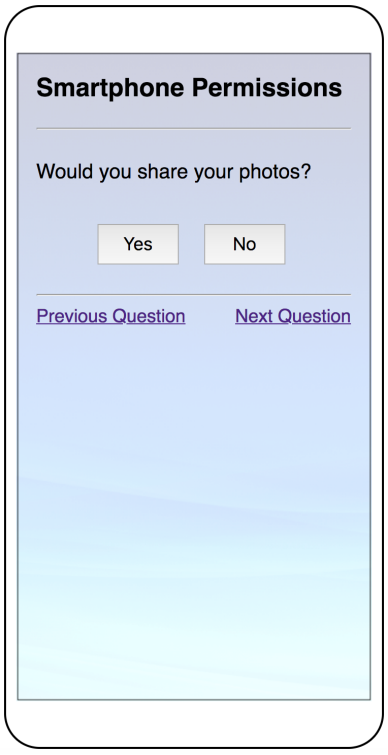
FIGURE 7.12: Interaction for picking a subprofile for the S set.



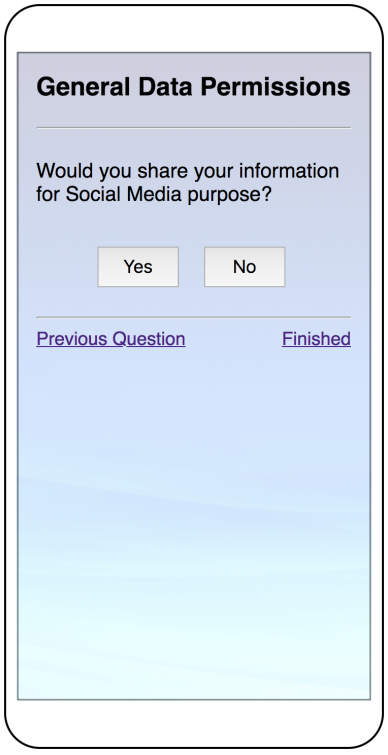
(A) A set.



(B) F set.

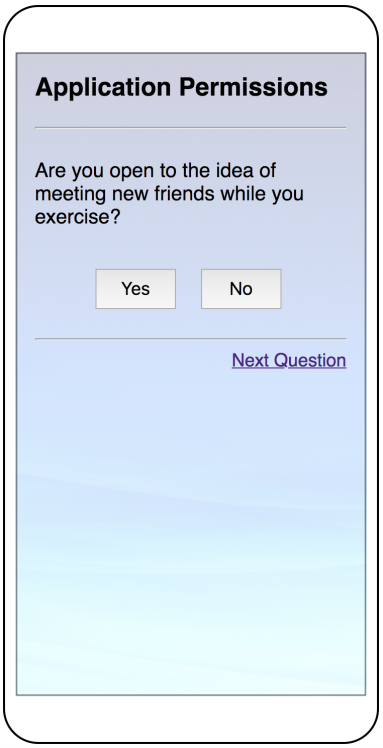


(C) S set.

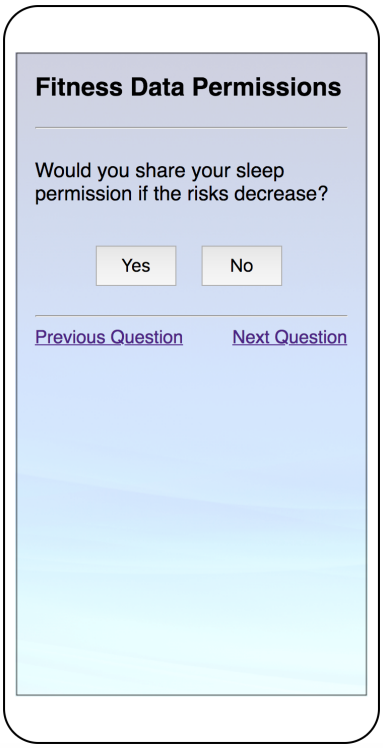


(D) G set.

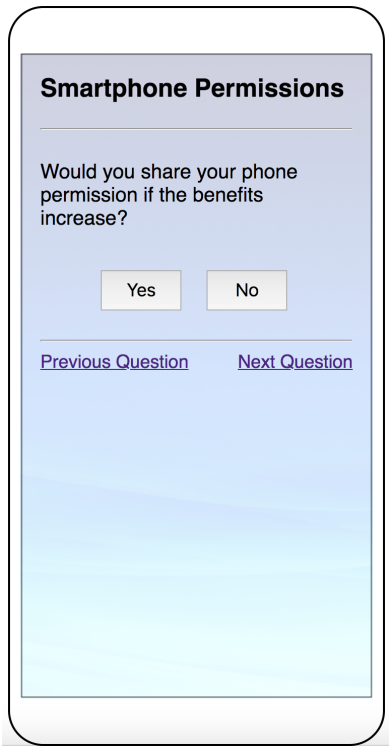
FIGURE 7.13: Direct Prediction questions.



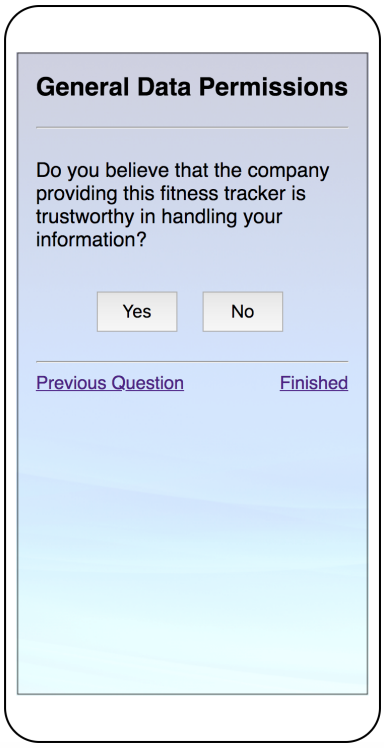
(A) A set.



(B) F set.



(C) S set.



(D) G set.

FIGURE 7.14: Indirect Prediction questions.

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%.

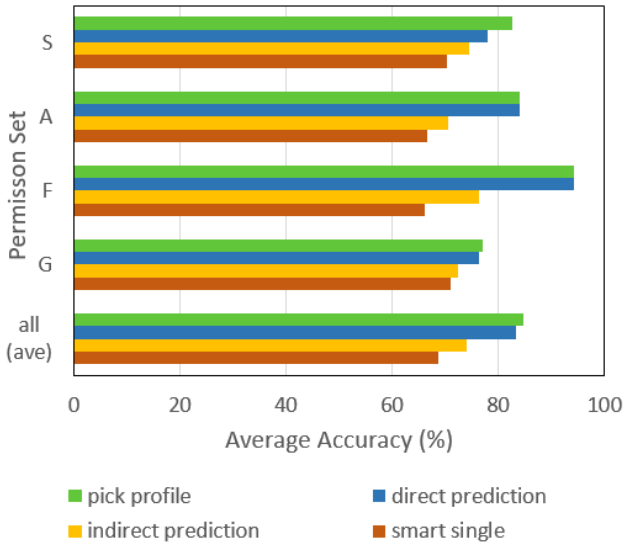


FIGURE 7.15: Average accuracies of the recommender strategies on the 30 users.

Finally, the *smart single default* approach uses only a single "profile", circumventing the need for classification. The default profile settings are shown in the 'full data' column of Figure 7.4. 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 and questionnaires used are provided in Appendix.

Chapter 8

Discussion and Conclusion

The PerNANDO framework aims to cover the literature gap on privacy protection specifically for IoT devices, which is in compliance with the GDPR. The framework consists of 2 main modules, namely, PDM, which manages and recommends user privacy settings, and AID-S, which computes risks for inference attacks. The framework uses the proposed PPIoT Ontology to model privacy preferences.

Our specific focus is on fitness wearables, given that this domain currently has most of the user information. Accordingly, we built a case study including the Fitbit fitness tracker, its app, and a third party app that can access and process the Fitbit data (i.e., API) to provide further services.

8.1 Discussion

RQ1 asks about how to protect IoT users from potential inference of their private data given their disclosed data. This has been addressed by AID-S which presents an approach to prevent the inference of users' private information from the available data released to third party applications.

An extensive analysis of the current information sharing model concerning third party applications has been investigated. By applying the framework to the case study, we showed the capability of AID-S to compute the risks of inference and recommend optimal settings to the users. This approach enables users to have enough information and be guided upon deciding which information they will share with the third parties, giving users a more informed consent.

We used the Bayesian Networks (BN) to discover inference risks and learn its structure and parameters using the dataset of the 49 users. Furthermore, the network has been integrated with inference studies from the literature and previous works.

BN is used since our goal is to create a graphical inference network and extend the current network of inferences in the literature. Also, BN is a probabilistic approach used in modeling risks (e.g., [56]), which is suitable for this study.

The main contribution of AID-S is to propose an approach to support users in managing privacy by revealing possible inference risks and also providing configuration options to limit such risk. In Android version 6.0 and above, the proposed approach becomes even more effective since users have the option to check/uncheck each permission needed by the third parties and AID-S can provide the recommendation of which information should not be released to them.

RQ2 asks about how to model user privacy preferences in the advent of heterogeneous personal IoT data. This has been addressed by proposing PPIoT Ontology that bridges the gap in semantic community on privacy preferences extended to the IoT context. More importantly, this ontology is compliant with the GDPR.

PPIoT Ontology can handle the complexity of heterogeneous personal IoT devices, and our PDM enables the expressive and fine-grained setting of privacy preferences. Our main contribution to the literature is the combination and extension of previous approaches for SW-based privacy management to cover the demands of the IoT domain.PDM also allows for negotiation on both the TP side and the user side, thereby balancing the privacy and utility of the service. Below, we summarize some of the improvements we made compared to existing work.

- The PPIoT Ontology improves upon PPO Ontology from [157] by providing a richer preference model that fits in the IoT paradigm, thereby allowing users to create fine-grained privacy preferences.

- The ontology proposed by Bodorik et al. [21] allows users to place regulations and conditions on factors based on the purpose of data recipient, usage and retention, disputes, remedy, and access control. This was included in the PPIoT extensions.

- Likewise, the PROACT Ontology [141] focuses on ubiquitous computing and includes a larger set of privacy concepts formalized in the notion of an "activity sphere". However, the abstract nature of this ontology makes it difficult to capture the complexity of the IoT paradigm. Arguably, our work captures a similar level of granularity.

- The ontology for privacy rules developed by Zhang et al. [214] addresses the privacy challenges of context-aware systems by defining a separate "data" class and a "condition" class. Our work adopts this approach, but also allows for *negotiation* of the conditions of each type of data between the user and the TP. We believe that this negotiation, in the context of an extensible SW Ontology, is crucial for the feasibility of privacy management in the context of IoT.

Finally, our PPIoT Ontology, as embedded in the interactive Privacy Preference Model, helps TPs meet the GDPR requirement of providing straightforward policy statements and requesting explicit consent.

RQ3 asks about how to aid users to set their privacy and provide them with suitable recommendation. This has been addressed by the PDM through creating privacy profiles and finding a way to match these profiles that best describes the user. Then, providing recommendation strategies to interact with the user. Thus, this research question has been broken into **RQ3.1**, **RQ3.2**, and **RQ3.3** as discussed extensively in Chapter 7.

We presented a data-driven approach to develop recommendation strategies for supporting users to set permissions on 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 applications in the Internet-of-Wearable-Things (IoWT), 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, in the light of the GDPR requirements.

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 four subsets of permissions (sets S, A, F and G), a total of 16 possible privacy profiles can be recommended to the user, which addresses **RQ3.1**.

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 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 is shared specifically. This confirms previous studies [109, 11] showing no significant interaction between these two parameters. Initial 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. All these predictors aim to address **RQ3.2**.

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 ontology: the data model vocabularies and the recommendation strategies will be used by the PDM to model the user privacy preferences and support privacy settings. The goal of providing interactive recommendation strategies was then achieved, which addressed **RQ3.3**.

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 systems that model and manage the user privacy permissions, like our PDM.

Our main contribution in this light is a process that can be used for the identification of privacy profiles and 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 influence.

As argued, though, this approach can also be applied to other IoT scenarios (e.g. household IoT, public IoT), Digital Humanities domains mentioned in Chapter 1 (e.g., Assisted Living, Digital hospitals, museum, tourism, , online learning, etc.) 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. [95]) in their own work. Applying to different digital humanities domain is possible, taking into account the PDM modeling of personal data used in the specific domain.

8.2 Limitations and Future Work

The thesis work clearly illustrate the contribution on managing the user's privacy preference in the proliferation of IoT devices and TP apps while satisfying the requirements of a lightweight and inter-operable approach. We discuss the main limitations of this work as follows.

No Policy Enforcement. An obvious weakness of our approach lies in the lack of control over the data once it has been disclosed to the TP. Third-party access can be allowed or restricted via the condition properties, but there are no guarantees that these properties are respected by the TP. Extensive data sharing among TPs can also result in additional inference risks. As such, our framework only applies to situations where TPs can be trusted and/or held accountable for their actions—but this is also true for traditional TP policy statements. This is why an apriori

protection for privacy inference is needed, where AID-S stands out. However, AID-S only protects the undisclosed data, not the data shared by user. The protection from malicious TPs are not addressed since it is part of the security-protected privacy as defined in Chapter 1.1.

Limited Negotiation Capability. Our proposal aims to automate the matching and negotiation between the TP policy statement and the user's preferences, resulting in a more transparent and controllable management of privacy permissions in the IoT context. This requires TPs to comply with the PPIoT Ontology. Note, though, that our approach uses "graceful degradation", where traditional privacy policies are automatically mapped to the PPIoT Ontology. The PDM can still operate in this case; only negotiation is not possible on the TPs part. Additionally, when the PDM automatic negotiation is not possible, it may result in a final manual decision that may be difficult for users. Our future work will address how the PDM recommendation can be phrased in a way that optimizes the user's confidence (cf. [132]).

Insufficient for Policy Replacement. Our work can help the TPs abide by the GDPR regulations by making their policies more transparent and controllable, and by allowing the TP to acquire explicit consent from the user as discussed in Chapter 6.2.1. However, we concur that PPM will not completely replace traditional policy statements, which are written in a certain way to provide legal protection. That said, we argue that our interactive privacy negotiation is a big step towards the newly implemented GDPR requirements.

Static Recommendation Strategy. For the PDM, one limitation is that the recommendation strategies are static. They do not update automatically based on new input. On the other hand, a dynamic recommender has some drawbacks. 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 is only once to set a user's privacy, 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 agree and give their permissions for their privacy preferences to be stored and processed.

As such, this thesis aims at studying which tracking data are viable for determining the right recommendation in a simplified manner. For future refinements, using dynamic techniques (e.g., Dynamic Bayesian Filtering, Kalman Filtering, PHD filtering, etc.) are feasible options that provide update steps to extend our static approach. Moreover, for future refinement, we plan also to combine direct recommendation and indirect recommendation, which are currently two different strategies that result from our study.

Generalizability. With regard to the PPIoT ontology, we provide an application-based evaluation through creating a simulation that is used to manage privacy preferences (offline) and by using it as the basis of the permission model which is evaluated by real users (online), as discussed in Chapter 6. Moreover, while its logical consistency has been evaluated using Jena Semantic Web reasoner, we are currently working on extending the use case scenarios in order to comprehensively evaluate its feasibility to model the user and TP privacy preferences in the IoWT.

User Evaluation on Recommendation Strategies. Another limitation of this thesis is the suitability of the recommendation strategies from the user's perspective. Though we have evaluated the system recommendation offline, an online user study validation must also be instantiated for the part on recommending strategies. 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. The user study should also evaluate the user control provided by the PDM.

Furthermore, the preliminary user evaluation results on AID-S should be extended. The interesting results show that there were more approvals after knowing the risks involved. These results must be verified and should be extended by checking if similar respondents (e.g., for the same fitness domain and/or on other IoT domains) provide similar behaviours.

Crowdsourced Dataset. Finally, we discuss a limitation of our dataset. The permission settings that we collected could be biased by the fact that the subjects knew it was a simulation of an app privacy setting. In order to reduce this possible effect, the interaction design and the user interface of the app were made very realistic and

we asked users to behave like when they usually install an app.

We also discuss the main open issues in this thesis. With regards to inference risks, one of the main open research issues is the completion of Inference matrix. iPshield [31] have used this method using mobile phones which we have then extended to the IoT domain. Given the different granularity and heterogeneity of data, computing all the possible inference risks would become a tedious and error-prone task. Finally, it will not be able to generalize into large-scale domains such as IoT.

Another open issue concerning privacy recommendation is the treatment of the privacy service provider as a third party. Although we are recommending user's privacy preferences, our framework should also be treated as a third party, which needs permission from the user. This research issue is general and also true for any privacy-preserving service, which is not tackled in this study. Most privacy schemes do not take this into account and this calls for more attention.

Concerning our PerNANDO framework, some open issues include the application of this framework to other IoT domains, which can also be done in our future work. Inside our framework, the Access Control work package (i.e., Policy Statement evaluation, and Authentication and Authorization) was not realized since these are heavily related to the security aspect. In this work we completely limit to the privacy point-of-view, as explained in Chapter 1.

Appendix A

Questionnaires and Accuracy Table

TABLE A.1: Study Questionnaire.

	Negotiability of privacy settings (Yes or No)
	Would you share the ff. data if the risks significantly increased? Would you share the ff. data if the benefits significantly decreased? Would you share the ff. data if the risks significantly decreased? Would you share the ff. data if the benefits significantly increased?
	Social behavior (7pt scale from Definitely Yes to Definitely No)
Social influence	If your friends exercise, does this influence you to exercise? If your online friends exercise, does this influence you to exercise?
Sociability	How often do you meet new friends while you exercise? choices: from *very often* to *never occured* Are you open to the idea of meeting new friends while you exercise?
	Exercise tendencies (7pt scale from Definitely Yes to Definitely No)
Exercise attitude	How physically healthy are you? choices: from *extremely healthy* to *extremely unhealthy* How important is exercise to you? choices: from *extremely important* to *completely unimportant* What do you most often do for exercise? choices: lift weights, walk/jog/run/hike, aerobics, dance, swim, pilates, other sports How often do you exercise? choices: from *almost daily* to *almost never* At what intensity do you work out? choices: from *very light* to *very heavy* Do you feel you get too much, the right amount, or too little exercise? choices: from *Much more than I want* to *Much less that i want* What is the main reason you exercise? choices: lose weight, for fun, get stronger/gain weight, manage stress, socialize w/ friends, maintain health, explore things
Healthy living expertise	I understand the diff. bet. 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.

	Privacy-related attitude (7pt scale from Completely Agree to Completely Disagree)
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.
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.

TABLE A.2: Table of Accuracies.

	Pick Profile	Smart Single	Direct Pred.	Privacy Attitude	Soc. Behavior	Neg.
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 %
Photos	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	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 %
Act. mins.	100.00 %	70.00 %	100.00 %	73.33 %	73.33 %	80.00 %
Cal. Act.	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 %
Heartrate	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 %
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 %

	Pick Profile	Smart Single	Direct Pred.	Privacy Attitude	Soc. Behavior	Neg.
G set						
SN Public	90.00 %	90.00 %	90.00 %	90.00 %	90.00 %	90.00 %
SN Friends	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 %

Bibliography

[1] Silvia Acid et al. “A comparison of learning algorithms for Bayesian networks: a case study based on data from an emergency medical service”. In: *Artificial intelligence in medicine* 30.3 (2004), pp. 215–232.

[2] Alessandro Acquisti, Laura Brandimarte, and George Loewenstein. “Privacy and human behavior in the age of information”. In: *Science* 347.6221 (2015), pp. 509–514.

[3] Alessandro Acquisti, Leslie K John, and George Loewenstein. “The impact of relative standards on the propensity to disclose”. In: *Journal of Marketing Research* 49.2 (2012), pp. 160–174.

[4] Yuvraj Agarwal and Malcolm Hall. “ProtectMyPrivacy: detecting and mitigating privacy leaks on iOS devices using crowdsourcing”. In: *Proceeding of the 11th annual international conference on Mobile systems, applications, and services*. ACM. 2013, pp. 97–110.

[5] Seyed Hossein Ahmadinejad, Philip WL Fong, and Reihaneh Safavi-Naini. “Privacy and utility of inference control mechanisms for social computing applications”. In: *Proceedings of the 11th ACM on Asia Conference on Computer and Communications Security*. ACM. 2016, pp. 829–840.

[6] Hazim Almuhimedi et al. “Your location has been shared 5,398 times!: A field study on mobile app privacy nudging”. In: *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. ACM. 2015, pp. 787–796.

[7] Claudio A Ardagna, Giovanni Livraga, and Pierangela Samarati. “Protecting privacy of user information in continuous location-based services”. In: *2012 IEEE 15th International Conference on Computational Science and Engineering*. IEEE. 2012, pp. 162–169.

[8] Mark Assad et al. “Giving users control over location privacy”. In: *Workshop on Ubicomp Privacy*. 2007.

[9] Jodie A Austin, Ian R Smith, and Amina Tariq. “The impact of closed-loop electronic medication management on time to first dose: a comparative study between paper and digital hospital environments”. In: *International Journal of Pharmacy Practice* (2018).

[10] Michael Backes et al. “Boxify: Full-fledged app sandboxing for stock android”. In: (2015).

[11] Paritosh Bahirat et al. “A Data-Driven Approach to Developing IoT Privacy-Setting Interfaces”. In: *23rd International Conference on Intelligent User Interfaces*. ACM. 2018, pp. 165–176.

[12] Adam Barth et al. “Privacy and contextual integrity: Framework and applications”. In: *Security and Privacy, 2006 IEEE Symposium on*. IEEE. 2006, 15–pp.

[13] Patricia Beatty et al. “P3P adoption on e-Commerce web sites: a survey and analysis”. In: *IEEE Internet Computing* 11.2 (2007).

[14] Victoria Bellotti and Abigail Sellen. “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*. Springer. 1993, pp. 77–92.

[15] Michael Benisch et al. “Capturing location-privacy preferences: quantifying accuracy and user-burden tradeoffs”. In: *Personal and Ubiquitous Computing* 15.7 (2011), pp. 679–694.

[16] David Benyon et al. “Presence and digital tourism”. In: *AI & society* 29.4 (2014), pp. 521–529.

[17] Alastair R Beresford et al. “Mockdroid: trading privacy for application functionality on smartphones”. In: *Proceedings of the 12th workshop on mobile computing systems and applications*. ACM. 2011, pp. 49–54.

[18] Greg Bigwood, F Ben Abdesslem, and Tristan Henderson. “Predicting location-sharing privacy preferences in social network applications”. In: *Proc. of AwareCast* 12 (2012), pp. 1–12.

[19] Igor Bilogrevic et al. "A machine-learning based approach to privacy-aware information-sharing in mobile social networks". In: *Pervasive and Mobile Computing* 25 (2016), pp. 125–142.

[20] Igor Bilogrevic et al. "Adaptive information-sharing for privacy-aware mobile social networks". In: *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*. ACM. 2013, pp. 657–666.

[21] Peter Bodorik, Dawn Jutla, and Mike Xuehai Wang. "Consistent privacy preferences (CPP): model, semantics, and properties". In: *Proceedings of the 2008 ACM symposium on Applied computing*. ACM. 2008, pp. 2368–2375.

[22] Stephen V Bowles et al. "The Use of Mindfulness and Acupuncture in the American Military". In: *Handbook of Military Psychology*. Springer, 2017, pp. 193–211.

[23] Janez Brank, Marko Grobelnik, and Dunja Mladenić. "A survey of ontology evaluation techniques". In: (2005).

[24] Ajay Brar and Judy Kay. *Privacy and security in ubiquitous personalized applications*. School of Information Technologies, University of Sydney, 2004.

[25] Rikus Bruwer, Hendrik Jacobus, and Riaan Rudman. "Web 3.0: governance, risks and safeguards". In: (2015).

[26] Letitia Burridge et al. "Person-centred care in a digital hospital: observations and perspectives from a specialist rehabilitation setting". In: *Australian Health Review* (2017).

[27] Zhipeng Cai et al. "Collective data-sanitization for preventing sensitive information inference attacks in social networks". In: *IEEE Transactions on Dependable and Secure Computing* 15.4 (2018), pp. 577–590.

[28] Francesca Carmagnola, Francesco Osborne, and Ilaria Torre. "Escaping the Big Brother: An empirical study on factors influencing identification and information leakage on the Web". In: *Journal of Information Science* 40.2 (2014), pp. 180–197.

[29] Francesca Carmagnola, Francesco Osborne, and Ilaria Torre. "User data discovery and aggregation: the CS-UDD algorithm". In: *Information Sciences* 270 (2014), pp. 41–72.

[30] Barbara Carminati et al. "Enhancing user control on personal data usage in internet of things ecosystems". In: *Services Computing (SCC), 2016 IEEE International Conference on*. IEEE. 2016, pp. 291–298.

[31] Supriyo Chakraborty et al. "ipShield: A Framework For Enforcing Context-Aware Privacy." In: *NSDI*. 2014, pp. 143–156.

[32] Anil Chaturvedi, Paul E Green, and J Douglas Caroll. "K-modes clustering". In: *Journal of Classification* 18.1 (2001), pp. 35–55.

[33] Amir Chaudhry et al. "Personal data: thinking inside the box". In: *Proceedings of the fifth decennial Aarhus conference on critical alternatives*. Aarhus University Press. 2015, pp. 29–32.

[34] Terence Chen et al. "On the effectiveness of obfuscation techniques in online social networks". In: *International Symposium on Privacy Enhancing Technologies Symposium*. Springer. 2014, pp. 42–62.

[35] Pern Hui Chia, Yusuke Yamamoto, and N Asokan. "Is this app safe?: a large scale study on application permissions and risk signals". In: *Proceedings of the 21st international conference on World Wide Web*. ACM. 2012, pp. 311–320.

[36] Eun Kyoung Choe et al. "Living in a glass house: a survey of private moments in the home". In: *Proceedings of the 13th international conference on Ubiquitous computing*. ACM. 2011, pp. 41–44.

[37] Delphine Christin et al. "A survey on privacy in mobile participatory sensing applications". In: *Journal of systems and software* 84.11 (2011), pp. 1928–1946.

[38] VNI Cisco Mobile. *Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2016–2021 White Paper*. 2017.

[39] Michael Compton et al. "The SSN ontology of the W3C semantic sensor network incubator group". In: *Web semantics: science, services and agents on the World Wide Web* 17 (2012), pp. 25–32.

[40] Andy Crabtree et al. "Enabling the new economic actor: data protection, the digital economy, and the Databox". In: *Personal and Ubiquitous Computing* 20.6 (2016), pp. 947–957.

[41] Justin Cranshaw, Jonathan Mugan, and Norman M Sadeh. "User-Controllable Learning of Location Privacy Policies With Gaussian Mixture Models". In: *AAAI*. 2011.

[42] George Danezis. "Inferring privacy policies for social networking services". In: *Proceedings of the 2nd ACM workshop on Security and artificial intelligence*. ACM. 2009, pp. 5–10.

[43] Dorothy E. Denning and Jan Schlorer. "Inference controls for statistical databases". In: *Computer* 7 (1983), pp. 69–82.

[44] Tamara Dinev and Paul Hart. "An Extended Privacy Calculus Model for E-Commerce Transactions". In: *Information Systems Research* 17.1 (Mar. 2006), pp. 61–80. DOI: 10.1287/isre.1060.0080. URL: http://isr.journal.informs.org/cgi/content/abstract/17/1/61.

[45] Cailing Dong, Hongxia Jin, and Bart P Knijnenburg. "Ppm: A privacy prediction model for online social networks". In: *International Conference on Social Informatics*. Springer. 2016, pp. 400–420.

[46] Yujie Dong et al. "A new method for measuring meal intake in humans via automated wrist motion tracking". In: *Applied psychophysiology and biofeedback* 37.3 (2012), pp. 205–215.

[47] Manuel Egele et al. "PiOS: Detecting Privacy Leaks in iOS Applications." In: *NDSS*. 2011, pp. 177–183.

[48] Mahmoud Elkhodr, Seyed Shahrestani, and Hon Cheung. "A semantic obfuscation technique for the Internet of Things". In: *Communications Workshops (ICC), 2014 IEEE International Conference on*. IEEE. 2014, pp. 448–453.

[49] Lavanya Elluri, Karuna Pande Joshi, et al. "A Knowledge Representation of Cloud Data controls for EU GDPR Compliance". In: *11th IEEE International Conference on Cloud Computing (CLOUD)*. 2018.

[50] Murat A Erdogdu and Nadia Fawaz. "Privacy-utility trade-off under continual observation." In: *ISIT*. 2015, pp. 1801–1805.

[51] European Commission. "A new era for data protection in the EU; What changes after May 2018". In: (2016), pp. 1–3. URL: https://ec.europa.eu/commission/sites/beta-political/files/data-protection-factsheet-changes_en.pdf.

[52] Loren E Falkenberg. "Employee fitness programs: Their impact on the employee and the organization". In: *Academy of Management Review* 12.3 (1987), pp. 511–522.

[53] Lujun Fang and Kristen LeFevre. "Privacy wizards for social networking sites". In: *Proceedings of the 19th international conference on World wide web*. ACM. 2010, pp. 351–360.

[54] Parvez Faruki et al. "Android security: a survey of issues, malware penetration, and defenses". In: *IEEE communications surveys & tutorials* 17.2 (2015), pp. 998–1022.

[55] Adrienne Porter Felt et al. "Android permissions: User attention, comprehension, and behavior". In: *Proceedings of the eighth symposium on usable privacy and security*. ACM. 2012, p. 3.

[56] Norman Fenton and Martin Neil. *Risk assessment and decision analysis with Bayesian networks*. Crc Press, 2012.

[57] Mario Frank et al. "Mining permission request patterns from android and facebook applications". In: *Data Mining (ICDM), 2012 IEEE 12th International Conference on*. IEEE. 2012, pp. 870–875.

[58] Nir Friedman et al. "Using Bayesian networks to analyze expression data". In: *Journal of computational biology* 7.3-4 (2000), pp. 601–620.

[59] Huiqing Fu et al. "A field study of run-time location access disclosures on android smartphones". In: *Proc. Usable Security (USEC)* 14 (2014), 10–pp.

[60] D Garson. "Creating simulated datasets". In: *Asheboro: North Carolina state University and G. David Garson and Statistical Associates Publishing* (2012).

[61] GDPR. "Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46". In: *Official Journal of the European Union (OJ)* 59.1-88 (2016), p. 294.

[62] Yuming Ge et al. "A comprehensive investigation of user privacy leakage to android applications". In: *Computer Communication and Networks (ICCCN), 2016 25th International Conference on*. IEEE. 2016, pp. 1–6.

[63] Robert Gellman. "Fair information practices: A basic history". In: (2017).

[64] Stylianos Gisdakis, Thanassis Giannetsos, and Panos Papadimitratos. "Android privacy C (R) ache: reading your external storage and sensors for fun and profit". In: *Proceedings of the 1st ACM workshop on privacy-aware mobile computing*. ACM. 2016, pp. 1–10.

[65] Asunción Gómez-Pérez. "Ontology evaluation". In: *Handbook on ontologies*. Springer, 2004, pp. 251–273.

[66] Google. "How people discover, use, and stay engaged with apps". In: (2016), pp. 1–15. URL: https://www.thinkwithgoogle.com/data/smartphone-users-discover-apps-browsing/.

[67] Ulrike Gretzel et al. "Smart tourism: foundations and developments". In: *Electronic Markets* 25.3 (2015), pp. 179–188.

[68] Junqi Guo et al. "Square-root unscented Kalman filtering-based localization and tracking in the Internet of Things". In: *Personal and ubiquitous computing* 18.4 (2014), pp. 987–996.

[69] Yang Guo, Hongbo Liu, and Yi Chai. "The embedding convergence of smart cities and tourism internet of things in China: An advance perspective". In: *Advances in Hospitality and Tourism Research (AHTR)* 2.1 (2014), pp. 54–69.

[70] Hamed Haddadi and Ian Brown. "Quantified self and the privacy challenge". In: *Technology Law Futures* 6 (2014).

[71] A Haller et al. "Semantic Sensor Network Ontology. W3C Recommendation". In: *World Wide Web Consortium, Oct* 19 (2017).

[72] Muhammad Haris, Hamed Haddadi, and Pan Hui. "Privacy leakage in mobile computing: Tools, methods, and characteristics". In: *arXiv preprint arXiv:1410.4978* (2014).

[73] Raymond Heatherly, Murat Kantarcioglu, and Bhavani Thuraisingham. "Preventing private information inference attacks on social networks". In: *IEEE Transactions on Knowledge and Data Engineering* 25.8 (2013), pp. 1849–1862.

[74] Stephan Heuser et al. "DroidAuditor: Forensic Analysis of Application-Layer Privilege Escalation Attacks on Android (Short Paper)". In: *International Conference on Financial Cryptography and Data Security*. Springer. 2016, pp. 260–268.

[75] Hlomani Hlomani and Deborah Stacey. "Approaches, methods, metrics, measures, and subjectivity in ontology evaluation: A survey". In: *Semantic Web Journal* 1.5 (2014), pp. 1–11.

[76] Dawn E Holmes and Lakhmi C Jain. "Introduction to Bayesian networks". In: *Innovations in Bayesian Networks*. Springer, 2008, pp. 1–5.

[77] Sherry Hsi and Holly Fait. "RFID enhances visitors' museum experience at the Exploratorium". In: *Communications of the ACM* 48.9 (2005), pp. 60–65.

[78] Yuh-Jong Hu and Jiun-Jan Yang. "A semantic privacy-preserving model for data sharing and integration". In: *Proceedings of the International Conference on Web Intelligence, Mining and Semantics*. ACM. 2011, p. 9.

[79] Zahid Iqbal et al. "Toward user-centric privacy-aware user profile ontology for future services". In: *Communication Theory, Reliability, and Quality of Service (CTRQ), 2010 Third International Conference on*. IEEE. 2010, pp. 249–254.

[80] Edwin T Jaynes. "Information theory and statistical mechanics". In: *Physical review* 106.4 (1957), p. 620.

[81] Carlos Jensen and Colin Potts. "Privacy policies as decision-making tools: an evaluation of online privacy notices". In: *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. ACM. 2004, pp. 471–478.

[82] Larry Johnson et al. *NMC horizon report: 2015 museum edition*. Tech. rep. The New Media Consortium, 2015.

[83] Lalana Kagal et al. "Authorization and privacy for semantic web services". In: *IEEE Intelligent Systems* 19.4 (2004), pp. 50–56.

[84] Svebor Karaman et al. "Personalized multimedia content delivery on an interactive table by passive observation of museum visitors". In: *Multimedia Tools and Applications* 75.7 (2016), pp. 3787–3811.

[85] Judy Kay, Bob Kummerfeld, and Piers Lauder. "Personis: a server for user models". In: *International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems*. Springer. 2002, pp. 203–212.

[86] Judy Kay and Bob Kummerfeld. "Scrutability, user control and privacy for distributed personalization". In: *Proceedings of the CHI2006 Workshop on Privacy-Enhanced Personalization*. 2006, pp. 21–22.

[87] Patrick Kelley et al. "A conundrum of permissions: installing applications on an android smartphone". In: *Financial cryptography and data security* (2012), pp. 68–79.

[88] Patrick Gage Kelley et al. "User-controllable learning of security and privacy policies". In: *Proceedings of the 1st ACM workshop on Workshop on AISec*. ACM. 2008, pp. 11–18.

[89] Daniel Kelly, Kevin Curran, and Brian Caulfield. "Automatic Prediction of Health Status Using Smartphone-Derived Behavior Profiles". In: *IEEE journal of biomedical and health informatics* 21.6 (2017), pp. 1750–1760.

[90] Rafiullah Khan et al. "Future internet: the internet of things architecture, possible applications and key challenges". In: *Frontiers of Information Technology (FIT), 2012 10th International Conference on*. IEEE. 2012, pp. 257–260.

[91] B. P. Knijnenburg. "A user-tailored approach to privacy decision support". English. Ph.D. Thesis. Irvine, CA: University of California, Irvine, 2015. URL: http://search.proquest.com/docview/1725139739/abstract.

[92] B. P. Knijnenburg, Alfred Kobsa, and Hongxia Jin. "Counteracting the Negative Effect of Form Auto-completion on the Privacy Calculus". In: *ICIS 2013 Proceedings*. Milan, Italy, 2013.

[93] B. P. Knijnenburg, Alfred Kobsa, and Hongxia Jin. "Dimensionality of information disclosure behavior". In: *International Journal of Human-Computer Studies* 71.12 (2013), pp. 1144–1162. ISSN: 1071-5819. DOI: 10.1016/j.ijhcs.2013.06.003. (Visited on 11/22/2013).

[94] Bart P Knijnenburg. "Information disclosure profiles for segmentation and recommendation". In: *SOUPS2014 Workshop on Privacy Personas and Segmentation*. 2014.

[95] Bart P Knijnenburg. "Privacy? I Can't Even! Making a Case for User-Tailored Privacy". In: *IEEE Security & Privacy* 15.4 (2017), pp. 62–67.

[96] Bart P Knijnenburg and Alfred Kobsa. "Helping users with information disclosure decisions: potential for adaptation". In: *Proceedings of the 2013 international conference on Intelligent user interfaces*. ACM. 2013, pp. 407–416.

[97] Bart P Knijnenburg and Alfred Kobsa. "Making decisions about privacy: information disclosure in context-aware recommender systems". In: *ACM Transactions on Interactive Intelligent Systems (TiiS)* 3.3 (2013), p. 20.

[98] Bart P Knijnenburg and Hongxia Jin. "The persuasive effect of privacy recommendations". In: *Twelfth Annual Workshop on HCI Research in MIS*. 2013.

[99] Alfred Kobsa. "Tailoring privacy to users' needs". In: *International Conference on User Modeling*. Springer. 2001, pp. 301–313.

[100] Trupti M Kodinariya and Prashant R Makwana. "Review on determining number of Cluster in K-Means Clustering". In: *International Journal* 1.6 (2013), pp. 90–95.

[101] Kevin B Korb and Ann E Nicholson. *Bayesian artificial intelligence*. CRC press, 2010.

[102] Noboru Koshizuka and Ken Sakamura. "Tokyo University Digital Museum." In: *Kyoto International Conference on Digital Libraries*. 2000, pp. 179–186.

[103] Lars Kulik. "Privacy for real-time location-based services". In: *SIGSPATIAL Special* 1.2 (2009), pp. 9–14.

[104] J Sathish Kumar and Dhiren R Patel. "A survey on internet of things: Security and privacy issues". In: *International Journal of Computer Applications* 90.11 (2014).

[105] Antonio Kung et al. "A Privacy Engineering Framework for the Internet of Things". In: *Data Protection and Privacy:(In) visibilities and Infrastructures*. Springer, 2017, pp. 163–202.

[106] Susan Landau. "What Was Samsung Thinking?" In: *IEEE Security & Privacy* 13.3 (2015), pp. 3–4.

[107] Scott Lederer, Jennifer Mankoff, and Anind K Dey. "Who wants to know what when? privacy preference determinants in ubiquitous computing". In: *CHI'03 extended abstracts on Human factors in computing systems*. ACM. 2003, pp. 724–725.

[108] Hosub Lee and Alfred Kobsa. "Privacy preference modeling and prediction in a simulated campuswide IoT environment". In: *Pervasive Computing and Communications (PerCom), 2017 IEEE International Conference on*. IEEE. 2017, pp. 276–285.

[109] Hosub Lee and Alfred Kobsa. "Understanding user privacy in Internet of Things environments". In: *Internet of Things (WF-IoT), 2016 IEEE 3rd World Forum on*. IEEE. 2016, pp. 407–412.

[110] Li Li et al. "Iccta: Detecting inter-component privacy leaks in android apps". In: *Proceedings of the 37th International Conference on Software Engineering-Volume 1*. IEEE Press. 2015, pp. 280–291.

[111] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. "t-closeness: Privacy beyond k-anonymity and l-diversity". In: *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*. IEEE. 2007, pp. 106–115.

[112] Yongchang Li et al. "A real-time EEG-based BCI system for attention recognition in ubiquitous environment". In: *Proceedings of 2011 international workshop on Ubiquitous affective awareness and intelligent interaction*. ACM. 2011, pp. 33–40.

[113] Daniel J Liebling and Sören Preibusch. "Privacy considerations for a pervasive eye tracking world". In: *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*. ACM. 2014, pp. 1169–1177.

[114] Jialiu Lin et al. "Expectation and purpose: understanding users' mental models of mobile app privacy through crowdsourcing". In: *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*. ACM. 2012, pp. 501–510.

[115] Jialiu Lin et al. "Modeling users' mobile app privacy preferences: Restoring usability in a sea of permission settings". In: (2014), pp. 199–212.

[116] Jie Lin et al. "A survey on internet of things: architecture, enabling technologies, security and privacy, and applications". In: *IEEE Internet of Things Journal* (2017).

[117] Bin Liu et al. "Follow my recommendations: A personalized privacy assistant for mobile app permissions". In: *Twelfth Symposium on Usable Privacy and Security*. 2016, pp. 26–41.

[118] Bin Liu, Jialiu Lin, and Norman Sadeh. "Reconciling mobile app privacy and usability on smartphones: Could user privacy profiles help?" In: *Proceedings of the 23rd international conference on World wide web*. ACM. 2014, pp. 201–212.

[119] Gi-Zen Liu and Gwo-Jen Hwang. "A key step to understanding paradigm shifts in e-learning: towards context-aware ubiquitous learning". In: *British Journal of Educational Technology* 41.2 (2010), E1–E9.

[120] Javier Lopez et al. "Evolving privacy: From sensors to the Internet of Things". In: *Future Generation Computer Systems* 75 (2017), pp. 46–57.

[121] Ashwin Machanavajjhala et al. "\ell-Diversity: Privacy Beyond\kappa-Anonymity". In: *null*. IEEE. 2006, p. 24.

[122] M. Madejski, M. Johnson, and S.M. Bellovin. "A study of privacy settings errors in an online social network". In: *Fourth International Workshop on Security and Social Networking*. SECSOC '12. Lugano, Switzerland, 2012, pp. 340–345. DOI: 10.1109/PerComW.2012.6197507.

[123] Naresh K Malhotra, Sung S Kim, and James Agarwal. "Internet users' information privacy concerns (IUIPC): The construct, the scale, and a causal model". In: *Information systems research* 15.4 (2004), pp. 336–355.

[124] Luciana Andreia Fondazzi Martimiano, Muller Roberto Pereira Goncalves, and Edson dos Santos Moreira. "An ontology for privacy policy management in ubiquitous environments". In: *Network Operations and Management Symposium, 2008. NOMS 2008. IEEE*. IEEE. 2008, pp. 947–950.

[125] Paul F Marty. "Museum websites and museum visitors: digital museum resources and their use". In: *Museum Management and Curatorship* 23.1 (2008), pp. 81–99.

[126] Carlo Maria Medaglia and Alexandru Serbanati. "An overview of privacy and security issues in the internet of things". In: *The Internet of Things*. Springer, 2010, pp. 389–395.

[127] Marci Meingast, Tanya Roosta, and Shankar Sastry. "Security and privacy issues with health care information technology". In: *Engineering in Medicine and Biology Society, 2006. EMBS'06. 28th Annual International Conference of the IEEE*. IEEE. 2006, pp. 5453–5458.

[128] Jonathan Mugan, Tarun Sharma, and Norman Sadeh. *Understandable learning of privacy preferences through default personas and suggestions*. 2011.

[129] Mark A Musen et al. "The protégé project: a look back and a look forward". In: *AI matters* 1.4 (2015), p. 4.

[130] Tanveer Mustafa and Karsten Sohr. "Understanding the implemented access control policy of Android system services with slicing and extended static checking". In: *International Journal of Information Security* 14.4 (2015), pp. 347–366.

[131] Alexios Mylonas, Marianthi Theoharidou, and Dimitris Gritzalis. "Assessing privacy risks in android: A user-centric approach". In: *International Workshop on Risk Assessment and Risk-driven Testing*. Springer. 2013, pp. 21–37.

[132] Moses Namara et al. "The Potential for User-Tailored Privacy on Facebook". In: *2018 IEEE Symposium on Privacy-Aware Computing (PAC)*. IEEE. 2018, pp. 31–42.

[133] Richard E Neapolitan et al. *Learning bayesian networks*. Vol. 38. Pearson Prentice Hall Upper Saddle River, NJ, 2004.

[134] Helen Nissenbaum. "Privacy as contextual integrity". In: *Wash. L. Rev.* 79 (2004), pp. 119–158.

[135] Patricia A. Norberg, Daniel R. Horne, and David A. Horne. "The Privacy Paradox: Personal Information Disclosure Intentions versus Behaviors". en. In: *Journal of consumer affairs* 41.1 (2007), pp. 100–126. ISSN: 1745-6606. DOI: 10.1111/j.1745-6606.2006.00070.x. URL: http://onlinelibrary.wiley.com/doi/10.1111/j.1745-6606.2006.00070.x/abstract (visited on 09/11/2013).

[136] Mahda Noura et al. "Concept Extraction from the Web of Things Knowledge Bases". In: *Proceedings of the International Conference WWW/Internet*. 2018.

[137] Natalya F Noy, Deborah L McGuinness, et al. *Ontology development 101: A guide to creating your first ontology*. 2001.

[138] Aleksandr Ometov et al. "Facilitating the delegation of use for private devices in the era of the internet of wearable things". In: *IEEE Internet of Things Journal* 4.4 (2017), pp. 843–854.

[139] Mahdi Ben Alaya oneM2M. "Webinar: Enhancing oneM2M platform with semantics, why does it matter?" In: (2016).

[140] Paolo Palatini and Stevo Julius. "Heart rate and the cardiovascular risk". In: *Journal of hypertension* 15.1 (1997), pp. 3–17.

[141] Ioannis Panagiotopoulos et al. "Proact: An ontology-based model of privacy policies in ambient intelligence environments". In: *Informatics (PCI), 2010 14th Panhellenic Conference on*. IEEE. 2010, pp. 124–129.

[142] Abhinav Parate et al. "Risq: Recognizing smoking gestures with inertial sensors on a wristband". In: *Proceedings of the 12th annual international conference on Mobile systems, applications, and services*. ACM. 2014, pp. 149–161.

[143] Tina R Patil and SS Sherekar. "Performance analysis of Naive Bayes and J48 classification algorithm for data classification". In: *International journal of computer science and applications* 6.2 (2013), pp. 256–261.

[144] Judea Pearl. "Bayesian networks: A model of self-activated memory for evidential reasoning". In: *Proceedings of the 7th Conference of the Cognitive Science Society, 1985*. 1985, pp. 329–334.

[145] Judea Pearl. "Causality: models, reasoning, and inference". In: *Econometric Theory* 19.675-685 (2003), p. 46.

[146] Charith Perera et al. "Privacy-knowledge modeling for the internet of things: A look back". In: *Computer* 49.12 (2016), pp. 60–68.

[147] Steven J Phillips, Robert P Anderson, and Robert E Schapire. "Maximum entropy modeling of species geographic distributions". In: *Ecological modelling* 190.3-4 (2006), pp. 231–259.

[148] Robert Porzel and Rainer Malaka. "A task-based approach for ontology evaluation". In: *ECAI Workshop on Ontology Learning and Population, Valencia, Spain*. Citeseer. 2004, pp. 1–6.

[149] M Prakash and G Singaravel. "An Analysis of Privacy Risks and Design Principles for Developing Countermeasures in Privacy Preserving Sensitive Data Publishing". In: *Journal of Theoretical & Applied Information Technology* 62.1 (2014).

[150] Dimitrios Rafailidis and Alexandros Nanopoulos. "Modeling users preference dynamics and side information in recommender systems". In: *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 46.6 (2016), pp. 782–792.

[151] Parisa Rashidi and Alex Mihailidis. "A survey on ambient-assisted living tools for older adults". In: *IEEE journal of biomedical and health informatics* 17.3 (2013), pp. 579–590.

[152] Sarunas J Raudys and Anil K. Jain. "Small sample size effects in statistical pattern recognition: Recommendations for practitioners". In: *IEEE Transactions on Pattern Analysis & Machine Intelligence* 3 (1991), pp. 252–264.

[153] Ramprasad Ravichandran et al. "Capturing social networking privacy preferences". In: *International Symposium on Privacy Enhancing Technologies Symposium*. Springer. 2009, pp. 1–18.

[154] Abbas Razaghpanah et al. "Haystack: In situ mobile traffic analysis in user space". In: *ArXiv e-prints* (2015).

[155] Jingjing Ren et al. "Recon: Revealing and controlling pii leaks in mobile network traffic". In: *Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services*. ACM. 2016, pp. 361–374.

[156] Owen Sacco and Alexandre Passant. "A Privacy Preference Manager for the Social Semantic Web." In: *SPIM*. 2011, pp. 42–53.

[157] Owen Sacco and John G Breslin. "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*. ACM. 2012, pp. 80–87.

[158] Norman Sadeh et al. "Understanding and capturing people's privacy policies in a mobile social networking application". In: *Personal and Ubiquitous Computing* 13.6 (2009), pp. 401–412.

[159] Torben Schinke, Niels Henze, and Susanne Boll. "Visualization of off-screen objects in mobile augmented reality". In: *Proceedings of the 12th international conference on Human computer interaction with mobile devices and services*. ACM. 2010, pp. 313–316.

[160] Marco Scutari. "Learning Bayesian networks with the bnlearn R package". In: *arXiv preprint arXiv:0908.3817* (2009).

[161] Marco Scutari and Jean-Baptiste Denis. *Bayesian networks: with examples in R*. Chapman and Hall/CRC, 2014.

[162] Tejal Shah et al. "Enhancing Automated Decision Support across Medical and Oral Health Domains with Semantic Web Technologies". In: *Proceedings of the 24th Australasian Conference on Information Systems* (Mar. 2014).

[163] Rob Shearer, Boris Motik, and Ian Horrocks. "HermiT: A Highly-Efficient OWL Reasoner." In: *OWLED*. Vol. 432. 2008, p. 91.

[164] Liping Shen, Minjuan Wang, and Ruimin Shen. "Affective e-learning: Using "emotional" data to improve learning in pervasive learning environment". In: *Journal of Educational Technology & Society* 12.2 (2009), pp. 176–189.

[165] Yulong Shen et al. "MicroThings: A Generic IoT Architecture for Flexible Data Aggregation and Scalable Service Cooperation". In: *IEEE Communications Magazine* 55.9 (2017), pp. 86–93.

[166] Cuiqi Si et al. "A group evolving-based framework with perturbations for link prediction". In: *Physica A: Statistical Mechanics and its Applications* 475 (2017), pp. 117–128.

[167] Evren Sirin and Bijan Parsia. "Pellet: An owl dl reasoner". In: *Proc. of the 2004 Description Logic Workshop (DL 2004)*. 2004, pp. 212–213.

[168] H Jeff Smith, Sandra J Milberg, and Sandra J Burke. "Information privacy: measuring individuals' concerns about organizational practices". In: *MIS quarterly* (1996), pp. 167–196.

[169] Daniel J Solove. "Introduction: Privacy self-management and the consent dilemma". In: *Harv. L. Rev.* 126 (2012), p. 1880.

[170] Sarah Spiekermann, Jens Grossklags, and Bettina Berendt. "Stated Privacy Preferences versus Actual Behaviour in EC Environments: a Reality Check". In: *WI-IF 2001: the 5th International Conference Wirtschaftsinformatik - 3rd Conference Information Systems in Finance*. Augsburg, Germany, 2001, pp. 129–148.

[171] Rudi Studer, V Richard Benjamins, and Dieter Fensel. "Knowledge engineering: principles and methods". In: *Data & knowledge engineering* 25.1-2 (1998), pp. 161–197.

[172] Chengwei Su et al. "Overview of Bayesian network approaches to model gene-environment interactions and cancer susceptibility". In: (2012).

[173] Yunchuan Sun et al. "Constructing the web of events from raw data in the web of things". In: *Mobile Information Systems* 10.1 (2014), pp. 105–125.

[174] Ali Sunyaev et al. "Availability and quality of mobile health app privacy policies". In: *Journal of the American Medical Informatics Association* 22.e1 (2014), e28–e33.

[175] Juliana Sutanto et al. "Addressing the Personalization-Privacy Paradox: An Empirical Assessment from a Field Experiment on Smartphone Users." In: *Mis Quarterly* 37.4 (2013).

[176] Latanya Sweeney. "k-anonymity: A model for protecting privacy". In: *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10.05 (2002), pp. 557–570.

[177] Ewelina Szczekocka et al. "Managing personal information: a telco perspective". In: *Proceedings of the 19th international innovations in clouds, internet and networks (ICIN)* (2016), pp. 1–8.

[178] Christopher Thompson et al. "When it's better to ask forgiveness than get permission: attribution mechanisms for smartphone resources". In: *Proceedings of the Ninth Symposium on Usable Privacy and Security*. ACM. 2013, 16–pp.

[179] Bhavani Thuraisingham et al. *Secure Data Provenance and Inference Control with Semantic Web*. Auerbach Publications, 2014.

[180] Ilaria Torre et al. "Fitness trackers and wearable devices: how to prevent inference risks?" In: *Proceedings of the 11th EAI International Conference on Body Area Networks*. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering). 2016, pp. 125–131.

[181] Ilaria Torre et al. "Supporting users to take informed decisions on privacy settings of personal devices". In: *Personal and Ubiquitous Computing* 22.2 (2017), pp. 345–364.

[182] Lynn Tsai et al. "Turtle Guard: Helping Android Users Apply Contextual Privacy Preferences". In: *Thirteenth Symposium on Usable Privacy and Security ({SOUPS} 2017)*. USENIX{ Association}. 2017, pp. 145–162.

[183] Dmitry Tsarkov and Ian Horrocks. "FaCT++ description logic reasoner: System description". In: *International joint conference on automated reasoning*. Springer. 2006, pp. 292–297.

[184] Anna M Turri, Ronn J Smith, and Steven W Kopp. "Privacy and RFID technology: a review of regulatory efforts". In: *Journal of Consumer Affairs* 51.2 (2017), pp. 329–354.

[185] Natan Uriely. "The tourist experience: Conceptual developments". In: *Annals of Tourism research* 32.1 (2005), pp. 199–216.

[186] Craig Van Slyke et al. "Concern for Information Privacy and Online Consumer Purchasing". In: *Journal of the Association for Information Systems* 7.1 (June 2006). URL: http://aisel.aisnet.org/jais/vol7/iss1/16 (visited on 03/15/2013).

[187] Michele Vescovi et al. "Building an eco-system of trusted services via user control and transparency on personal data". In: *IFIP International Conference on Trust Management*. Springer. 2015, pp. 240–250.

[188] Carmen Ruiz Vicente et al. "Location-related privacy in geo-social networks". In: *IEEE Internet Computing* 15.3 (2011), pp. 20–27.

[189] Johanna Virkki and Liquan Chen. "Personal perspectives: Individual privacy in the IoT". In: *Advances in Internet of Things* 3.02 (2013), p. 21.

[190] Denny Vrandečić. "Ontology evaluation". In: *Handbook on ontologies*. Springer, 2009, pp. 293–313.

[191] Yang Wang, Yun Huang, and Claudia Louis. "Towards a framework for privacy-aware mobile crowdsourcing". In: *2013 International Conference on Social Computing*. IEEE. 2013, pp. 454–459.

[192] Jason Watson, Heather Richter Lipford, and Andrew Besmer. "Mapping user preference to privacy default settings". In: *ACM Transactions on Computer-Human Interaction (TOCHI)* 22.6 (2015), p. 32.

[193] Alan F Westin. "Privacy and freedom". In: *Washington and Lee Law Review* 25.1 (1968), p. 166.

[194] Primal Wijesekera et al. "Android Permissions Remystified: A Field Study on Contextual Integrity." In: *USENIX Security Symposium*. 2015, pp. 499–514.

[195] Primal Wijesekera et al. "The feasibility of dynamically granted permissions: Aligning mobile privacy with user preferences". In: *Security and Privacy (SP), 2017 IEEE Symposium on*. IEEE. 2017, pp. 1077–1093.

[196] Shomir Wilson et al. "Privacy manipulation and acclimation in a location sharing application". In: *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*. ACM. 2013, pp. 549–558.

[197] Pamela Wisniewski, Bart P Knijnenburg, and H Richter Lipford. "Profiling facebook users privacy behaviors". In: *SOUPS2014 Workshop on Privacy Personas and Segmentation*. 2014.

[198] Pamela J Wisniewski, Bart P Knijnenburg, and Heather Richter Lipford. "Making privacy personal: Profiling social network users to inform privacy education and nudging". In: *International Journal of Human-Computer Studies* 98 (2017), pp. 95–108.

[199] Ian H Witten et al. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

[200] Philip A Wolf et al. "Probability of stroke: a risk profile from the Framingham Study." In: *Stroke* 22.3 (1991), pp. 312–318.

[201] Anthony Wood et al. "ALARM-NET: Wireless sensor networks for assisted-living and residential monitoring". In: *University of Virginia Computer Science Department Technical Report* 2 (2006), p. 17.

[202] Jingzheng Wu et al. "POSTER: biTheft: stealing your secrets by bidirectional covert channel communication with zero-permission android application". In: *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*. ACM. 2015, pp. 1690–1692.

[203] Le Wu et al. "Modeling users' preferences and social links in social networking services: a joint-evolving perspective". In: *Thirtieth AAAI Conference on Artificial Intelligence*. 2016.

[204] Jierui Xie, Bart Piet Knijnenburg, and Hongxia Jin. "Location sharing privacy preference: analysis and personalized recommendation". In: *Proceedings of the 19th international conference on Intelligent User Interfaces*. ACM. 2014, pp. 189–198.

[205] Heng Xu et al. "Examining the formation of individual's privacy concerns: Toward an integrative view". In: *ICIS 2008 proceedings* (2008), p. 6.

[206] Heng Xu et al. "Measuring mobile users' concerns for information privacy". In: (2012).

[207] Heng Xu et al. "The personalization privacy paradox: An exploratory study of decision making process for location-aware marketing". In: *Decision Support Systems* 51.1 (Apr. 2011). ACM ID: 1943793, pp. 42–52. ISSN: 0167-9236. DOI: 10.1016/j.dss.2010.11.017. (Visited on 02/28/2011).

[208] Feng Yan et al. "Distributed autonomous online learning: Regrets and intrinsic privacy-preserving properties". In: *IEEE Transactions on Knowledge and Data Engineering* 25.11 (2013), pp. 2483–2493.

[209] Tong Yan, Yachao Lu, and Nan Zhang. "Privacy disclosure from wearable devices". In: *Proceedings of the 2015 Workshop on Privacy-Aware Mobile Computing*. ACM. 2015, pp. 13–18.

[210] Ibrar Yaqoob et al. "Internet of things architecture: Recent advances, taxonomy, requirements, and open challenges". In: *IEEE wireless communications* 24.3 (2017), pp. 10–16.

[211] Weiwei Yin et al. "A tree-like Bayesian structure learning algorithm for small-sample datasets from complex biological model systems". In: *BMC systems biology* 9.1 (2015), p. 49.

[212] YIN Yuehong et al. "The internet of things in healthcare: An overview". In: *Journal of Industrial Information Integration* 1 (2016), pp. 3–13.

[213] Roberto Yus and Eduardo Mena. "Mobile Endpoints: Accessing Dynamic Information from Mobile Devices." In: *International Semantic Web Conference (Posters & Demos)*. 2015.

[214] Ni Zhang and Chris Todd. "Developing a privacy ontology for privacy control in context-aware systems". In: *Dept. of Electronic & Electrical Eng., Univ. College London* (2006).

[215] Weizhe Zhang et al. "Android platform-based individual privacy information protection system". In: *Personal and Ubiquitous Computing* 20.6 (2016), pp. 875–884.

[216] Yuchen Zhao, Juan Ye, and Tristan Henderson. "Privacy-aware location privacy preference recommendations". In: *Proceedings of the 11th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering). 2014, pp. 120–129.

[217] Xiaoyong Zhou et al. "Identity, location, disease and more: Inferring your secrets from android public resources". In: *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*. ACM. 2013, pp. 1017–1028.

[218] Guy Zyskind, Oz Nathan, et al. "Decentralizing privacy: Using blockchain to protect personal data". In: *Security and Privacy Workshops (SPW), 2015 IEEE*. IEEE. 2015, pp. 180–184.