Alma Mater Studiorum Università di Bologna
Archivio istituzionale della ricerca

Where You Go Matters: A Study on the Privacy Implications of Continuous Location Tracking

(Article begins on next page)

07 January 2025

Where You Go Matters: A Study on the Privacy Implications of Continuous Location Tracking

BENJAMIN BARON, University College London, UK

MIRCO MUSOLESI, University College London, UK, The Alan Turing Institute, UK, and University of Bologna, Italy

Data gathered from smartphones enables service providers to infer a wide range of personal information about their users, such as their traits, their personality, and their demographics. This personal information can be made available to third parties, such as advertisers, sometimes unbeknownst to the users. Leveraging location information, advertisers can serve ads micro-targeted to users based on the places they visited. Understanding the types of information that can be extracted from location data and implications in terms of user privacy is of critical importance.

In this context, we conducted an extensive *in-the-wild* research study to shed light on the range of personal information that can be inferred from the places visited by users, as well as privacy sensitivity of the personal information. To this end, we developed TrackingAdvisor, a mobile application that continuously collects user location and extracts personal information from it. The app also provides an interface to give feedback about the relevance of the personal information inferred from location data and its corresponding privacy sensitivity. Our findings show that, while some personal information such as social activities is not considered private, other information such as health, religious belief, ethnicity, political opinions, and socio-economic status is considered private by the participants of the study. This study paves the way to the design of privacy-preserving systems that provide contextual recommendations and explanations to help users further protect their privacy by making them aware of the consequences of sharing their personal data.

CCS Concepts: • **Human-centered computing** → **Smartphones**; *Empirical studies in collaborative and social computing*; • **Security and privacy** → *Social aspects of security and privacy.*

Additional Key Words and Phrases: Location tracking, personal information inference, self check-in mobile systems

1 INTRODUCTION

An increasing number of location-based services continuously track the locations of users, often without their knowledge [6, 8, 10, 12, 14, 17, 20, 33, 37, 53, 60, 61, 68]. This data is very valuable as it reveals personal information about the users [57]. Applications, such as Google Assistant (initially launched as Google Now [33]) or Siri, and third parties (*e.g.*, advertisers) can leverage it to present marketing information or content tailored to the user's interests and personal preferences.

Golbeck *et al.* and Chitkara *et al.* both showed that users are largely unaware of the privacy implications of some permissions that users grant to applications and services, in particular access to location information [12, 29]. In fact, the inferred personal information is diverse and includes sensitive data such as the user's place of residence, interests, and religion [40, 41, 55]. A plethora of applications and studies have shown that developers and third parties can leverage location data with machine learning techniques to infer user personality and demographics from digital traces shared on social networks [13, 27, 33, 35, 39, 43, 44, 47, 52]. However, there has been limited interest to date in understanding the range of personal information that can be inferred from location data and how these inferences might violate users' privacy [13, 27, 29, 40, 67]. Indeed, existing research only focuses on inferring a small subset of places, generally limited to home and work locations [38, 53], but do not evaluate the privacy implications of the personal information one can extract from the places users visit on the user preferences.

Proc. ACM Interact. Mob. Wearable Ubiquitous Technol., Vol. 4, No. 4, Article 169. Publication date: December 2020.

169

In this paper, we present a research study that aims at understanding the privacy sensitivity of the information that can be extracted from location data. In the current context where applications are *Through this study, we propose to show what personal attributes that can be automatically inferred from continuously-collected location data and understand what is their perceived privacy sensitivity.* The goal is to understand the categories of personal information that are privacy-sensitive when shared with third-party entities such as advertisers. The results of the study might help in understanding the implications of the technologies and they might be used for deriving general guidelines for application developers and system designers.

In order to obtain ground truth data, we have carried out an *in-the-wild* research study that collects real-world location traces and gathers user feedback about the personal information extracted from the places they visit. For this study, we have developed an iOS application, TrackingAdvisor[1] which continuously collects the users' location data. The application automatically extracts location data, infers the personal information associated with the places visited by the users and presents it to them. This relies on a general framework for the extraction of personal information from location data that we have developed. Even if the proof of concept implementation is based on specific datasets, including OpenStreetMap [34], Foursquare [1], and DBPedia [5], it can be generalizable and applied to a variety of application scenarios. TrackingAdvisor has been specifically designed to collect ground truth data and users' feedback on the different elements that are automatically extracted and inferred. We asked for the users' feedback regarding the places they visited in order to collect ground truth data. This information is generally accessible to third parties such as advertisers. User feedback is used to evaluate both the importance of the personal information and its perceived privacy sensitivity. Analysis of user feedback sheds a light on the nuanced privacy sensitivity perceptions of some personal information and helps determine the implications for the design of privacy-preserving systems. We summarize our main contributions as follows:

- We have conducted an *in-the-wild* research study with 69 participants over the course of three months. Throughout the study, we continuously tracked the participants' location and collected ground truth data about the places they visited, the personal information associated with these places, and their privacy sensitivity. Unlike most of the user studies that passively collect data and makes inferences about users' behavior and characteristics, our study actively gathers feedback from them, allowing us to collect their actual privacy preferences.
- We have performed an in-depth analysis of the perceived privacy sensitivity of the personal information that can be extracted from the places they visit. Moreover, the research study draws implications for the design of future personalized privacy-preserving systems that take into account the users' privacy preferences.
- We provide a general methodology for understanding privacy implications of location information collected by means of mobile technologies, which can be applied to a variety of studies involving data of personal nature in different populations and contexts.

The paper is structured as follows. In Section 2, we present the related work and in Section 3, we detail the implementation of the application, including the mechanisms used to automatically extract users' interests from the places they visit. We present the results of the study in Section 4. Finally, we discuss the results and the implications of the study in Section 5, before summarizing the key contributions of this work in Section 6.

---

[1]The real name of the application has been replaced by this name for the double-blind review process. We will change it back for the camera-ready version of this paper in case it is accepted for publication. We will also add the link for downloading the application from the Apple App Store.

2 RELATED WORK

In this section, we review the related work in two key areas, namely the techniques that infer personal information about users from the data they generate and studies that explore the privacy concerns of users with location-based services.

2.1 Personal Information Inference

A large body of work has investigated the possibility of inferring traits and personal information about individuals using social media platforms such as Facebook or Twitter [13, 16, 27, 35, 39, 43, 44, 47, 49, 52]. In the following, we summarize the most representative works that infer personal information about users from their interactions with digital systems.

Social media. Interactions between users and brands within social media can reveal the users' interests and preferences. Using Facebook *likes* collected by the myPersonality dataset, Kosinski *et al.* inferred the traits and personality of the users [39]. In more detail, the authors used personality questionnaires and collected the Facebook *likes* of several thousands of participants from the US in order to analyze their predictive power for a wide collection of personal and sensitive traits such as sexual orientation, ethnic origin, political views, religion, personality, intelligence, and substance use, as well as other demographics attributes, including age, gender, and relationship status. The authors used linear and logistic regression models to infer and predict individual traits and attributes with high accuracy from the Facebook *likes*.

Search engines. Queries made on search engines such as Google Search, AOL, and Bing have proven to reveal user interests and preferences [31, 49, 54, 58]. These are mainly used for the purpose of "personalized search" as well as to increase the click-through rate of ads shown in sponsored searches. For instance, Bi *et al.* leveraged the myPersonality dataset to predict the traits and personality of users from their search queries made on a search engine [8]. The authors successfully predicted the age and gender with an accuracy of 74% and 80%, respectively.

Physical interactions. Interactions in the "physical" world can also allow external parties to infer personal information including the users' traits, personality, and demographics, whether they buy groceries at their local supermarket [61], use their smartphone [68] or use their credit card [20].

Location information. Personal attributes can also be inferred from location information, available in various forms [13, 19, 30, 66, 67]. Gonzalez *et al.* showed that anonymized mobile phone location traces surface reproducible individual mobility patterns, which can then be used to predict the routine and next movements of individuals [30]. Also using mobile phone location traces, De Montjoye *et al.* showed that individuals can be easily singled out from a large-scale dataset using just a few datapoints [19]. These studies show the vulnerability of individuals' privacy with location traces. Chorley *et al.* proposed a method to infer personality traits from users' places of interest, visited in addition to the home and work locations [13]. The authors collected the "Big Five" personality traits as well the Foursquare check-ins of participants and found significant correlations between the mobility patterns of the users and different traits, including Conscientiousness, Openness, and Neuroticism. Zhong *et al.* developed a so-called "location-to-profile" framework that aims at understanding user demographic attributes including marital status, gender, age, and education background from the patterns of their check-ins in a location-based social network [67]. Specifically, the authors have leveraged the profile information given by the users when registering to the social network as a basis for a supervised learning approach that learns the attributes from spatio-temporal check-in patterns and place features, including categories, and user reviews. Zhong *et al.* proposed a framework that leverages mobile data collected from smartphones to predict demographic attributes such as gender, job type, marital status and age [66]. The framework relies on a supervised method to learn and predict with a good accuracy the attributes from spatio-temporal features extracted from the mobility traces. Contrary to our work, the authors did not consider the places that the users of the dataset

have visited as a basis for the prediction and do not study the privacy implications of the predictions. Finally, several works have leveraged mobility data collected from phones to monitor and infer specific physical and mental health conditions (see for example [11, 48]). We did not consider this type of inference in our study as this paper focuses on the personal information one can extract from visits to places, but indeed the analysis of the full mobility traces might reveal additional sensitive information about the users.

Contrary to the works presented above, we aim at understanding the perceived importance and sensitivity of information inferred from the mobility data. Indeed, the personal information derived from location data only gives a partial picture of the personal characteristics of an individual. Although this picture could be completed using other sources of data such as interactions in social networks, search logs or purchase history, we wanted to show that location data may surface important personal information that can be perceived as sensitive by the individuals and can be accessed by third parties without the user to be fully aware of the implications. An investigation about the potential fusion of different sources of information is part of our future research agenda.

2.2 Studies on the Privacy Concerns with Location-Based Services

There has been several studies on privacy concerns and location-based services [6, 10, 12, 14, 17, 37, 53, 57, 60]. In the following, we describe the most representative works in this area.

Location tracking-services. Barkhuus and Dey conducted one of the first research study with the goal of understand the privacy concerns of users when accessing location-based services [6]. Junglas and Waston addressed the same problem in a later study [37]. In both of them, the authors distinguish two types of location-based services as previously defined by Snekkenes [56]: (1) location-tracking services that allow third parties to collect the location of the user's phone and (2) position-aware services that only allow the phone itself to know the user's location. While the participants of the studies perceived the location tracking-services as more useful in general than position-aware services, they also rated the former services as more intrusive, as their location would be shared with other parties instead of remaining on their phone. Using a bidding mechanism, Danezis *et al.* evaluated the privacy value of location data in a student population with a study that periodically queries the location of the participants' phone [18]. The authors found that students value their location privacy even more in commercial settings. Staiano *et al.* found similar trends in the study they conducted a few years later to evaluate the sensitivity of personal information belonging to four broad categories: communication, applications, location and media with varying levels of granularity [57]. The authors show through a reverse auction mechanism that the location information reported as distance travelled, places visited and GPS positions is the most valuable and sensitive category of information. In our work, we are specifically interested in better understanding the perception of the privacy of location data generated as part of location tracking-services.

Privacy preferences. Khalil and Connelly [38], Cvrcek *et al.* [17], Brush *et al.* [10], and Toch *et al.* [59] have explored how individuals value their location data by conducting different user studies. The authors try to understand the users' concerns about the continuous collection and sharing of their location information to third parties such as family and friends, colleagues, private companies, and academic institutions. In these studies, the authors found that most of the participants were willing to share their location information to these third parties in an anonymous manner or to trade their information to use location-based services for free or in exchange for useful services. In particular, Khalil and Connelly studies the privacy preferences of their participants according to their willingness to disclose personal information including location, activity, company, and current availability to different entities such a family, friends, and work colleagues. The authors studies the sharing rate of home and work locations to the different entities and found that privacy is more desirable at work than at work and are less likely to share their location information with their colleagues and boss than with their significant other. In the same vein and more recently, Shih *et al.* examined the privacy preferences of users when it comes to share their personal context (*i.e.*, where they are, who they are with, and what they are doing) with third parties,

namely other mobile applications installed on their phone [53]. The authors showed that participants' decision to share data is affected by the sensitivity of the data and the purpose for collecting the data. They became more willing to disclose their context, even for private location, when the usage of the data (purpose) is either clearly stated or missing. They further showed that the purpose of the use of the location and its context is important to the users and often not clearly indicated by the permission system of the phone. Vague or non-explanatory purposes reminded the users of the trade-off with the privacy risk and the benefits of sharing the data, which made them unwilling to share it. This suggests some improvements to the mobile phone permission system to make the implications of sharing the data clearer with respect to privacy.

Privacy permissions. A series of projects have explored the impact of privacy permissions that users grant to applications [2, 4, 12, 29]. In particular, Chitkara *et al.* have developed ProtectMyPrivacy, an Android applica-tion that aims to infer the context around data accesses of applications [12]. More specifically, ProtectMyPrivacy provides privacy-related suggestions to users to help them decide whether to deny or allow the access of the application to sensitive data items such as location, contacts, and identifiers. The authors performed stack-trace analysis to determine the purpose of the access for each data item, in particular whether it is used by the appli-cation itself or by third-party libraries (*e.g.*, for further monitoring or advertisement purposes). Via a user study, the authors showed that adding transparency to the purpose of the data accesses increases the users' concern and caution with regard to the behaviors of third-party services. Finally, Golbeck *et al.* studied how users are concerned with privacy and whether they know what information they share with third parties when using Facebook applications [29]. The authors showed that the participants are concerned for their privacy, in par-ticular with the personal information that Facebook applications can access. Through their study, the authors showed that a large percentage of participants were unaware that apps could access certain data about them and it was necessary to educate them so that they get a better understanding of how their data is shared, with whom, and what are the privacy implications of that sharing.

Limitations of the current studies. We have summarized the studies we have reviewed in Table 1. These studies treat the location data as a whole, without making any distinctions among the different types of personal information that can be inferred from the location data. With the increasing number of applications that rely on third-party libraries and the poor knowledge of privacy permissions [12, 29], we argue that we need to take into account the various types of personal information that can be extracted from location data in the context of sharing it to third parties. Further, since the relative perception in terms of privacy-sensitive is different, information about a user's favorite park does not reveal as much information as the information about a user's regular church or hospital. The studies we have reviewed do not take into account the inferences that can be made from the location data. This can limit the privacy awareness of users when it comes to sharing their location with third parties such as advertisers. In our work, we propose to study the privacy expectations and their implications by considering the type and extent of personal information that can be inferred from location data. To the best of our knowledge, our work is the first that focuses on these aspects.

Table 1. Summary of the related work with respect to our work on location privacy and personal information inference.

Related work	Type of study	Type of personal information	Granularity	Target audience
Barkhuus and Dey [6]	Journal with prefilled questions ($N = 16$)	Perceived usefulness of position-aware and location tracking services	Places and GPS location	Private and public settings, friends
Junglas and Waston [37]	Task analysis with pre-defined tasks ($N = 58$)	Perceived usefulness and ease of use of position-aware and location tracking services	Places and GPS location	Acquaintances and public
Danezis et al. [18]	Online survey on mobile phone ($N = 74$)	Perceived privacy value of location data with reverse auction mechanism	GPS location	Acquaintances
Staiano et al. [57]	Survey on mobile phone ($N = 60$)	Perceived privacy value of communication, applications, location, and media data with reverse auction	Individual, processed, and aggregated	Banks, government, insurance companies, telcos, and user
Khalil and Connelly [38]	ESM time diary with Palm PDA ($N = 20$)	Willingness to share location, activity, company, and availability	Location GPS and places (home and work)	Significant other, family member, friend, colleague, boss and unknown
Cvrcek et al. [17]	Online questionnaire in five EU countries ($N = 1200$)	Perceived privacy value of location data using bidding mechanisms	GPS location	Academic and commercial
Brush et al. [10]	GPS logger ($N = 32$)	Monetary value of location data	GPS location with name or anonymous and with or without obfuscation method	Public, corporate, and academic
Toch et al. [59]	ESM mobile application ($N = 28$)	Willingness to share location data	Location GPS with semantic tags and time	Friends and acquaintances
Shih et al. [53]	ESM mobile application with semi-structured personalized surveys ($N = 34$)	Privacy sensitivity to share location data with other applications	Places (home, work, leisure, transport)	Unknown, commercial, user
Almuhimedi et al. [4]	ESM mobile Android application launcher for AppOps ($N = 23$)	Privacy permissions of applications with nudges and reports	Application access to personal data (location, call logs, contacts, calendar)	Application (developers and third-party libraries)
Agarwal et al. [2]	iOS mobile application ($N = 90, 621$)	Privacy permissions of applications with recommendations and reports	Application access to personal data (location, identifier, contacts, music library)	Application (developers and third-party libraries)
Chitkara et al. [12]	ESM Android mobile application ($N = 1, 321$)	Privacy permissions with access to personal data (location, contacts, and identifiers)	Application and third-party access to data	Application (developers and third-party libraries)
Golbeck et al. [29]	Questionnaire and Facebook application ($N = 120$)	Personal data access by a Facebook application	Facebook basic information	Application (developers and third-party libraries)
Our work	ESM iOS mobile application ($N = 69$)	Perceived importance and privacy sensitivity of personal information categories	Personal information categories	Commercial (ads)

Fig. 1. Overview of the two components and their interactions: (1) the mobile application TrackingAdvisor and (2) the backend pipeline.

3 DESCRIPTION OF THE RESEARCH STUDY

In this section, we describe the *in-the-wild* research study we conducted and the components of the application, including the overall system for location tracking and personal information extraction we developed as part of it. We first discuss the motivations of the research study. We then provide an overview of the architecture of the overall system and we present the implementation of TrackingAdvisor, the mobile application that collects user location data and presents the data back to the user. We also present the implementation of the personal information inference component and the associated review process that is showed to the users.

3.1 Motivation for the Research Study

Successive research studies have shown that the different permissions required by applications to access various components such as the contact list, the phone's microphone or the phone's location are poorly understood by users [4, 25, 53]. As a result, users often grant the applications access to the requested permission without knowing *what* data will be made accessible to the application, *which* entity has access to it and *how* the data can be processed, in particular how its processing can infringe on their own privacy.

As suggested by numerous works, location information can reveal a wide range of personal and sensitive information about a user [13, 27, 67]. However, these works have mainly focused on extracting information such as the users' home and work places, and to the best of our knowledge, no work has tried to infer an extensive amount of personal information from location information in order to understand their privacy sensitivity. In this context, we believe that it is important to raise user awareness by showing them the personal information items that can be inferred from the continuous tracking of their location. We also think that it is important to understand whether users consider the sharing of personal information with third parties as sensitive.

3.2 Overview of the Research Study

Given the objectives of the proposed investigation, we designed and carried out an *in-the-wild* research study to collect location data and obtain feedback about the importance and privacy sensitivity of personal information

(a) Timeline where users can confirm the visits detected at places.

(b) Timeline showing the visits confirmed by the users.

(c) Place edit screen where the user can change the time or place of the visit.

(d) Place search screen with place suggestions in the vicinity of the center of the map.
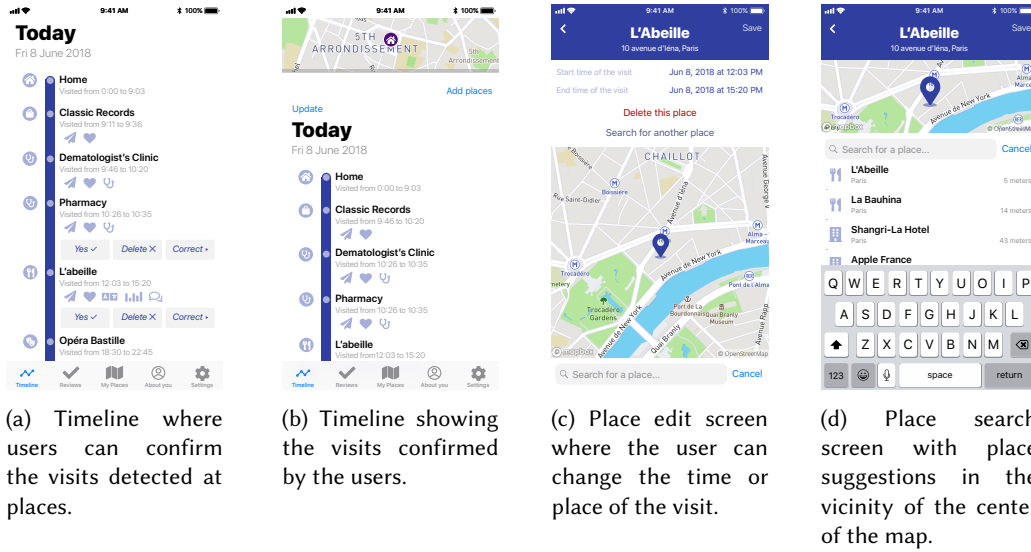
Fig. 2. Screenshots of the timeline (2a and 2b) and the place edit screen (2c and 2d) captured from the application TrackingAdvisor.

inferred automatically based on the places visited directly from the users. To this end, we developed TrackingAdvisor, a native iOS application in charge of automatically collecting the locations from the participants' phones and providing them with a way to give feedback on the automatic inferences performed by the system. As such, this study relies on the Experience Sampling Method (ESM) [15], as the application acts as a daily diary where users can report their feedback and thoughts while guaranteeing that the users are familiar with the places they review, since they have visited them. The main rationale for an ESM study was for the participants to use the application as a self-tracking diary, as they would with similar applications and not feel the pressure of a controlled study.

We built a backend system in charge of processing the location traces collected by the mobile application. We represent the two main components in Figure 1.[2] In the following, we briefly discuss the functionalities of the mobile application and the backend, together with the main interactions among them.

We have developed an energy-efficient location tracking system that detects places users visit using the iOS Core Location framework[3]. The location tracking trades off the accuracy of the locations with the frequency of the location updates. Periodically, the phone sends the collected locations to the backend, which then extracts places and personal information as described in the following paragraph. We provide more details about our location tracking system in Appendix A.

The backend pipeline is in charge of (1) collecting the user location traces from the users' phones, (2) processing the traces to extract the significant stay points, (3) matching the stay point to actual visited places, (4) inferring the corresponding personal information, and (5) collecting the user feedback. The backend server has dedicated API endpoints to receive location traces from the users' phones. Upon receipt of the location traces, they are saved in a geographic database with indexes on the geo-coordinates of the location points.

---

[2]For privacy reasons, the screenshots shown in the figures of this paper have been created for illustrative purposes and do not correspond to any specific user.

[3]Core Location. https://developer.apple.com/documentation/corelocation.

Using the location traces, this pipeline extracts the significant stay points. A stay point corresponds to a stop the user made at a place, typically ranging from a few minutes (*e.g.*, stopping by the convenience store to get milk) to several hours (*e.g.*, at home or at work). The pipeline then matches the significant stay points the place that the user most likely visited. We use open-access place databases, including OpenStreetMap [34] and Foursquare [1] to reverse-geocode coordinates into places. For full reproducibility, we provide a detailed description of the system and details about the databases in Appendix B.

With the visited places, the pipeline then infers a set of personal information using the place metadata available in open-access place databases. We classify the broad spectrum of personal information into a taxonomy with distinct personal information categories such that it covers a specific range of an individual's identity and personality, including religious beliefs, political opinions, sexuality and gender, and topics of interest. We then match the places and their categories in the place database (*e.g.*, church, university, coffee shop) with the relevant personal information (*e.g.*, religious beliefs, occupation, and topics of interest, respectively). We discuss the implementation of this component in Section 3.3.

Together, the components of the pipeline allow us to extract the places visited by the users and infer personal information about them. The mobile application TrackingAdvisor presents an overview of the visits made each day by the user through a timeline back to the user, from the first visit of the day at the top to the most recent visit at the bottom of the timeline. We present screenshots of the application in Figures 2a and 2b. The user can navigate from one day to another by swiping the timeline to the left for the previous day and to the right for the next day. For each visit, we present the name of the place and the start and end times of each visit. We also present interactive buttons that the user can tap to either confirm the visit, reject (when it is a false positive), or correct it (when it has been matched to the wrong place). This feedback is exploited to collect information about the places that the users have actually visited and to improve the learning and personalization of the place matching algorithm. In particular, we will give more weight to the visits that have been confirmed by the user when matching them with a detected stay point (refer to Section 3.3).

When the visit has been matched to the wrong place, the user can replace it by the correct place. We depict the place search screen in Figures 2c and 2d. By default (*i.e.*, when no text is searched), the search algorithm returns the 20 top places in the 200-meter radius around the center of the map displayed on the screen. If no places are found, the radius is incrementally increased until one is found.

3.3 Personal Information Inference

As mentioned in the previous section, we infer a set of personal information about each place from their description and metadata from different place databases, including OpenStreetMap [34], Foursquare [1], and DBPedia [5]. A personal information item relates to an individual by revealing one or several aspects of their identity and personality. While a personal information item can directly identify individuals with their name or address, it can also indirectly identify them by characterizing their racial or ethnicity affinity, political opinions, religious or philosophical beliefs, sex life, and sexual orientations [21, 22, 45]. Such personal information is useful for third parties, such as advertisers to target specific individuals or a group of individuals, as part of micro-targeted advertising. While there are no established classifications for personal information, we structure them in broad categories.

We link the places across place databases using the location and place name information as described in Appendix B.2. A place is further characterized by a set of categories (*e.g.*, office, university, coffee shop, hospital) that reveals personal information. For instance, users who regularly visit a hospital could either be working at the hospital or be patients receiving a treatment for an illness (Health). Similarly, a user who visits a veterinarian most likely owns a pet (Ownership) and is interested in animals (Topics of interest). Users who regularly visit restaurants of specific world cuisines most likely might have affinities with the related ethnicity (Ethnic
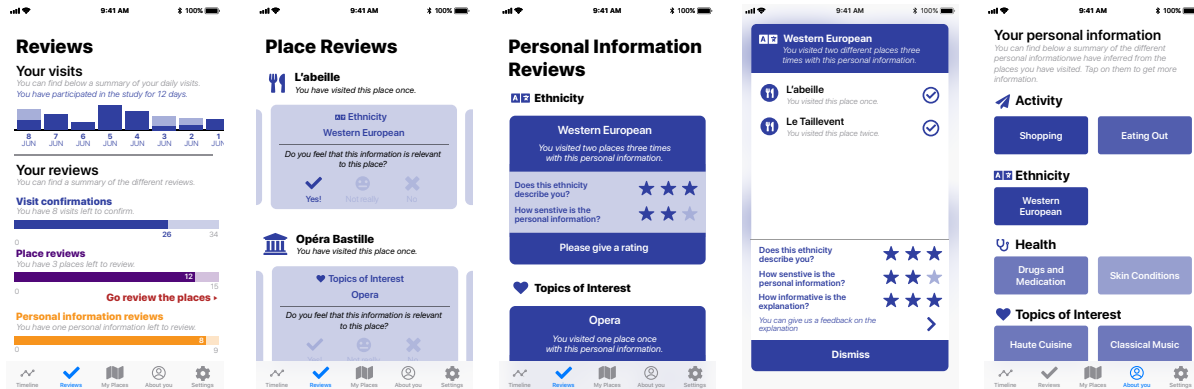
(a) Screen showing the summary of the reviews that the user has given and the pending ones.

(b) Review of the personal information associated to a place, that were automatically extracted.

(c) Personal information review of those aggregated from the places that the user has visited.

(d) Detail of the Figure 3c showing the places visited associated with the personal information.

(e) Capture of the screen "About you" showing a summary of the personal information reviewed by the user.

Fig. 3. Screenshots of the different review forms and summary captured from the application TrackingAdvisor.

affinities). With the exception of the Topics of interest category, we manually match the personal information categories with high-level place categories. We analyze the privacy sensitivity of the different personal information categories in Section 4.4. The different personal information categories are listed in Table 4 presented in the Appendix.

Through our mobile application, we offer the possibility to add new personal information to the places. Once a user has added a personal information to a place, it is added to a pool of personal information to be validated and then added to the general pool of personal information associated to a place. This gives an opportunity to crowdsource and curate the personal information associated to the places. A newly added personal information can also be relevant to the places that are part of the same chain (*e.g.*, the personal information Coffee is relevant to the all places of the chain Starbucks). We take this fact into account by populating the personal information across different places within the same chain.

3.4 TrackingAdvisor Personal Information Review Process and Summary

In our mobile application TrackingAdvisor, we designed UI elements that enable users to give explicit feedback. In particular, once the user has confirmed a visit at a place as shown in Figure 2a, the application invites the user to first review the relevance of the personal information items with respect to the places visited. Once the relevance of the personal information items for the visited place have been rated, the user can evaluate the importance and the privacy of each personal information item. As we show in Figure 3a with an instance of our app TrackingAdvisor, we proceed in two steps with two distinct review pools: (1) place reviews and (2) personal information reviews. This two step process ensures that users only review the relevant personal information of the places that they have visited. Thus, the system discards the wrong places and the irrelevant personal information items so they are not shown to the user for review.

Place Personal Information Review Task. In the first step, we ask the user to review the personal information items associated to the places *for which they have confirmed their visit*, as depicted in Figure 3b. This guarantees that the user will not review the personal information items for places they did not visit. To this end, the user

can rate the relevance of the personal information item with the following scale: *Yes!*, *Maybe*, and *No*. Users select *Yes!* if the personal information is very relevant to the place they visited; they select *Maybe* if the personal information item is somewhat relevant to the place, but not necessarily to all their visits at the place (*e.g.*, if user usually goes to a drugstore to get a snack and not a pill for headaches, the user will select a high relevance to the personal information item Snacks and a low relevance to the personal information item Drugs and medication); and a user selects *No* if the personal information item is not relevant at all for the place. This step is important, as with the rated personal information items we collect ground truth data about the different places and the relevant corresponding personal information items.

Personal Information Review Task. In the second step, we ask the user to review the *importance* of the personal information items aggregated from all the places visited with respect to themselves. As shown in Figures 3c and 3d, we aggregate the personal information items extracted from each visited place that have confirmed by the user as part of the previous review process. As a result, one personal information item is the aggregation of its different instances in all of the places that the user has visited. For each aggregated personal information, we ask the user to say whether it is important to themselves. For instance, in the case of a topic of interest, we ask how the personal information item relates to the user and give a 3-point scale with three stars from "Low" to "High". This allows us to understand whether the aggregated information we have inferred about the user is important. We also ask the user to rate the sensitivity of the personal information item on a similar 3-point scale. Since privacy preferences of users are contextual, as shown by Shih *et al.* [53], we provided the users with a global usage scenario that defines the overall context of the study in which they choose to share the personal information items. In particular, we asked the users to give their privacy preferences when it comes to sharing the personal information item automatically inferred with third parties, including advertisers. This allows us to understand which personal information items are considered as sensitive by the users, that is whether they would be willing to share and disclose them to third parties. We also provide an explanation regarding the aggregated personal information that shows a list of all the places with an instance of the personal information item and the number of visits were made at each of the places. We further ask the user to rate the explanation provided with another 3-point scale in order to evaluate the interpretability of the explanations. As we show in Figure 3d, the user also has the possibility of writing a free-text comment about the explanations in order to provide a detailed feedback.

Personal Information Summary. We summarize the personal information reviewed items by the user in a screen titled "About you" represented in Figure 3e in an ordered list that shows, for each personal information category, the aggregated personal information items. We rank each personal information displayed in the list according to their importance for the user, such that personal information items aggregated from a large number of places visited several times will be more relevant than the personal information aggregated from fewer places, which were rarely visited.

4 RESULTS OF THE RESEARCH STUDY

In this section, we present a comprehensive evaluation of the main algorithmic aspects of the system for personal information inference from location and a summary of the results of our study.

4.1 Description of the Research Study

Recruitment of the Participants. TrackingAdvisor was published on the Apple App Store and advertised on traditional social media platforms, including Twitter, Facebook, relevant mailing lists, as well as on websites dedicated to promote research studies. The application has been installed by 81 participants, 69 of whom have completed the research study for a minimum duration of two weeks. Note that the users of the application chose to participate in the study without any financial compensation. Among these participants, 38% were women,

48% were men, and 14% chose not to disclose their gender; 52% were over 35, 39% were less than 35, and 9% chose not to disclose their age. The top five countries where participants lived are: United Kingdom (23%), USA (22%), France (14%), Germany (12%), and Australia (6%).

Ensuring Privacy Compliance. In order to allow TrackingAdvisor to continuously collect the user location in the background, the users must give an explicit permission to collect the location, as required by iOS when requesting always-on location services.[4] Moreover, in order to comply with the latest privacy regulations, on its first launch, TrackingAdvisor shows a consent form and a privacy policy that both detail what information is collected, how the information is used, and whether the information collected will be shared to other parties. The user has to explicitly agree to both the consent form and the privacy policy before being able to use the application and participate in the user study. This process ensures that the user goes through a multi-level user agreement and is completely aware of the type of information collected by the application. The data was transmitted securely from the phone to our backend servers hosted within our institution's premises. The users had the possibility to opt out from the study at any given time, in which case the data associated to them would be deleted automatically if they wished it. Participants could further ask for their data through the form integrated in the application.

It is worth noting that the application has received the ethical approval and the data protection compliance, including all procedures and material, from the Research Ethics Committee at University College London.

4.2 Collected Dataset

We analyzed the data of the 69 users who participated in the research study for a minimum of two weeks. The dataset associated to these users comprises 205,143 individual location points, 2,467 unique places visited, 12,786 total visits at places, 19,578 unique personal information associated to places, and 4,867 unique personal information aggregated from the instances of the visited places. In the following, we give some insights we were able to derive from our dataset.

(a) Feedback received from the users for the different review tasks part of the review workflow.

(b) Cumulative distribution function of the number of personal information items per place visited by the users.
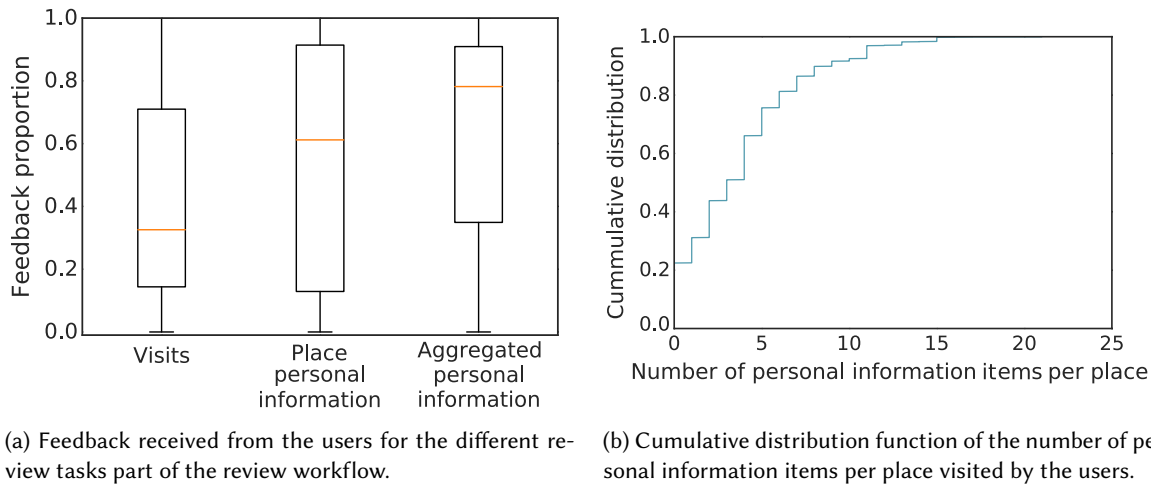
Fig. 4. Key statistics about user feedback.

---

[4]Apple website. *About privacy and Location Services in iOS 8 and later.* Available at the following link: https://support.apple.com/en-gb/HT203033.

Feedback Given by the Users. Through our mobile application TrackingAdvisor, we show the information that our backend has extracted from the user locations, including the places visited and the personal information inferred. We further ask the user to confirm the visits at places and the relevance of the personal information. As we presented in Figure 1, the feedback given by the users follows a review workflow with the following successive review tasks: (1) the users confirm the correctness of the detected visits, (2) the users rate the relevance of the personal information for each new place whose visit has been confirmed, and (3) the users rate the importance of the personal information items aggregated from the places they visited. We summarize key statistic of the feedback we have received from the users following this workflow during the study in Figure 4a. The visit confirmation is the task that has received the least feedback with an average of 38.7% feedback received. This figure is explained by the way the application has been designed: since the application shows the current day with the visits to confirm, users have little incentive to go back the previous days to confirm the visits. Despite this, we used different techniques to nudge the users to give feedback with daily reminders via push notifications. The place personal information and aggregated personal information tasks have a higher average feedback rate of 62.3% and 79.5%, respectively. We stress that the personal information items of the first review task are only associated to the places with confirmed visits, and the percentage derives from the total number of personal information associated to those places. Similarly, the personal information items of the second review task are aggregated from the personal information items rated as relevant in the preview review task. The lower feedback rate of the first review task (place personal information review) can be explained by its repetitive nature, as some personal information items would re-occur several times for different places (*e.g.*, the social activities or the activity associated to the place). Overall, each place has an average number of 3.78 personal information items rated as relevant by the users, as shown in Figure 4b.

4.3 Personal Information Relevance Evaluation

In this section, we present an evaluation of the different personal information items rated by the users during the study. In particular, we report the relevance with respect to the places associated with the personal information. Recall from Section 3.3 that the users were presented the list of the places for which they confirmed their visit and, for each place, a list of related personal information items. As shown in Figure 3b, users had to rate the relevance of each personal information item with respect to the place on a 3-point scale (*Yes*, *Maybe*, *No*). We plot the feedback results using diverging stacked bar charts, a common visualization to represent varying opinions [51].

In the following, we present the user feedback we received using diverging stacked bar charts. We also provide a statistical analysis of the level of agreement of the feedback we collected from the users using the Kappa test [26]. Fleiss' κ has been established as one of the most popular measures to assess the reliability of agreement among different raters when assigning categorical, nominal ratings, such as those we collected on the importance of the personal information items and their privacy sensitivity. We have filtered the categories by discarding those with less than five unique users who gave feedback. κ is equal to 1 if all the raters are in complete agreement, while a negative value for κ denotes a poor agreement among the users. Table 2 lists the level of agreements that are commonly used depending on the value of κ [46].

We present the relevance of the personal information items inferred from places as a diverging stacked bar chart in Figure 5. For each personal information category, the chart shows the percentage of the different feedback aggregated from all the users that evaluates the relevance of the inferred personal information: (1) "Not relevant" corresponds to the button *No*; (2) "Somewhat relevant" corresponds to the button *Maybe*; and (3) "Very relevant" corresponds to the button *Yes*. The number of personal information items rated by the users is presented in the figure (with the number of unique users who rated them in parenthesis). We did not display the personal information categories with less than 10 ratings received. The majority of the personal information that was automatically inferred was relevant to the places visited by the users, for 75% of the personal information

Table 2. Interpretation of the κ value according to [46].

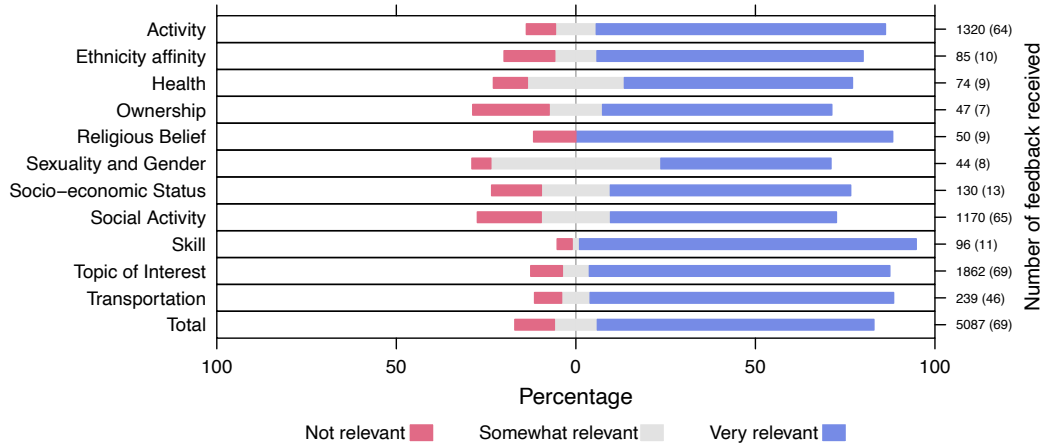| κ	Interpretation
0 − .20	No agreement
.21 − .39	Minimal agreement
.40 − .59	Weak agreement
.60 − .79	Moderate agreement
.80 − 0.90	Strong agreement
.80 − 1	Almost perfect agreement

Fig. 5. Relevance of the inferred personal information. The numbers on the right-hand side are the numbers of feedback received and in parenthesis the number of unique users that provided feedback.

items overall. 11% of the personal information items were considered not relevant and 14% of the personal information items were considered somewhat relevant. This demonstrates the accuracy of our personal inference methodology detailed in Section 3.3.

4.4 Personal Information Importance

Perceived Importance of Aggregated Personal Information Items. We evaluate the perceived importance of personal information items extracted and aggregated from the places users visited. The UI interface presented in Figure 3c allowed users to rate the importance using a 3-point scale, ranging from *Not important* to *Very important.*

Figure 6 shows the importance of the aggregated personal information items. Overall, the users consider 14% of the personal information as not important, 37% as somewhat important, and 49% as very relevant to themselves. These results contrast with those we presented with Figure 5, where users rated the relevance of the personal information with respect to the places. While some places may be associated to a set of personal information, the personal information items have different levels of importance for the users. For instance, without loss of generality, in the case of Health, while most of the personal information is somewhat or very relevant to the places (92%), a significant proportion of the personal information is not perceived as important for the users (31%). This can be explained by considering the fact that a wide range of personal information is inferred about
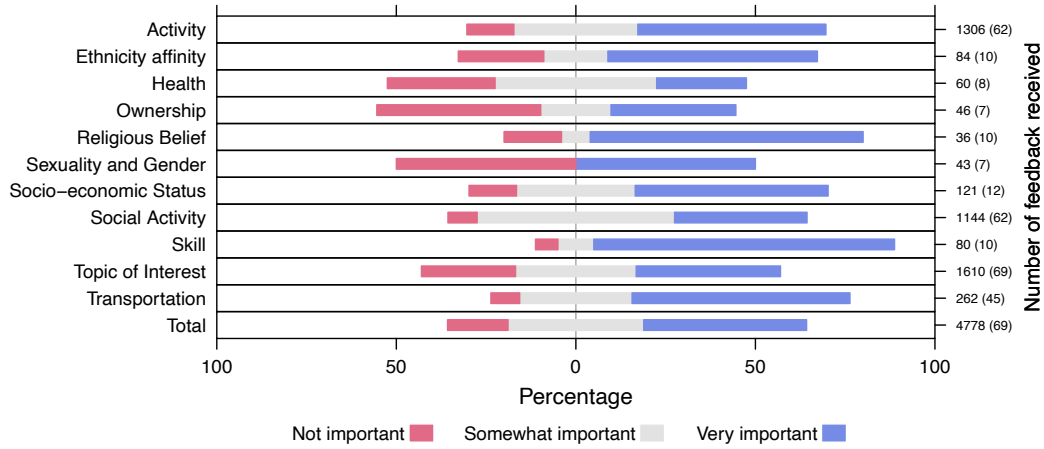
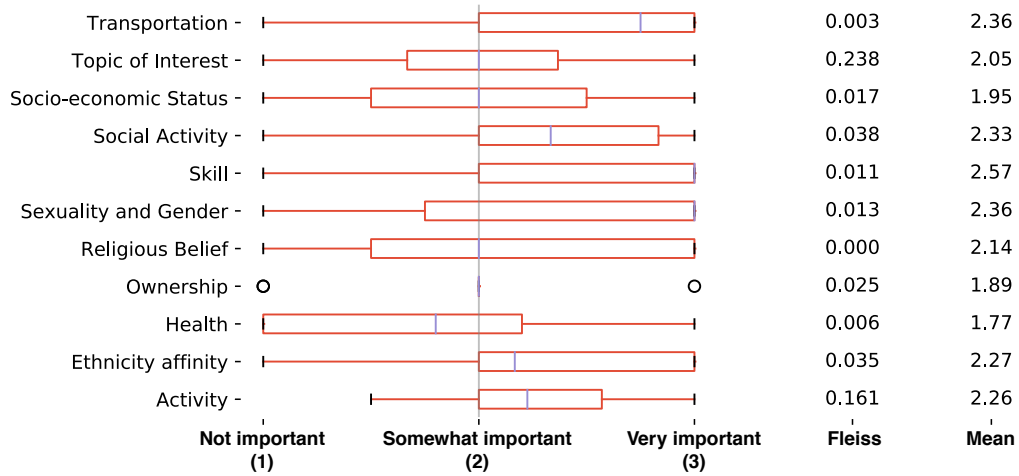Fig. 6. Perceived importance of the aggregated personal information items.

Fig. 7. Box plot of the user feedback distribution and Kappa test results for the perceived importance of the aggregated personal information categories with the corresponding Fleiss' kappa to measure the level of agreement and average mean of user feedback.

the place, and only a few personal information correspond to the profile of the user and the purpose of their visit to the place.

In Figure 7, we provide more details on the levels of agreements of the users for each personal information category. Some categories, such as Health and Religious Belief have no agreement on their perceived importance ($\kappa = 0.006$ and $\kappa = 0$, respectively) while the Topic of Interest category has minimal agreement on its perceived importance ($\kappa = 0.238$, respectively) as somewhat important. In addition to the agreement measure provided by Fleiss' Kappa, in the boxplots we can see that there is a high variability of opinions across the users. This

denotes different levels of perceived importance for the personal information categories, in particular for Topic of Interest.

Perceived Importance of the Topics of Interest. We will now analyze whether the perceived importance of the different topics of interest, as Topic of Interest is the category that has the highest rate of user's feedback.
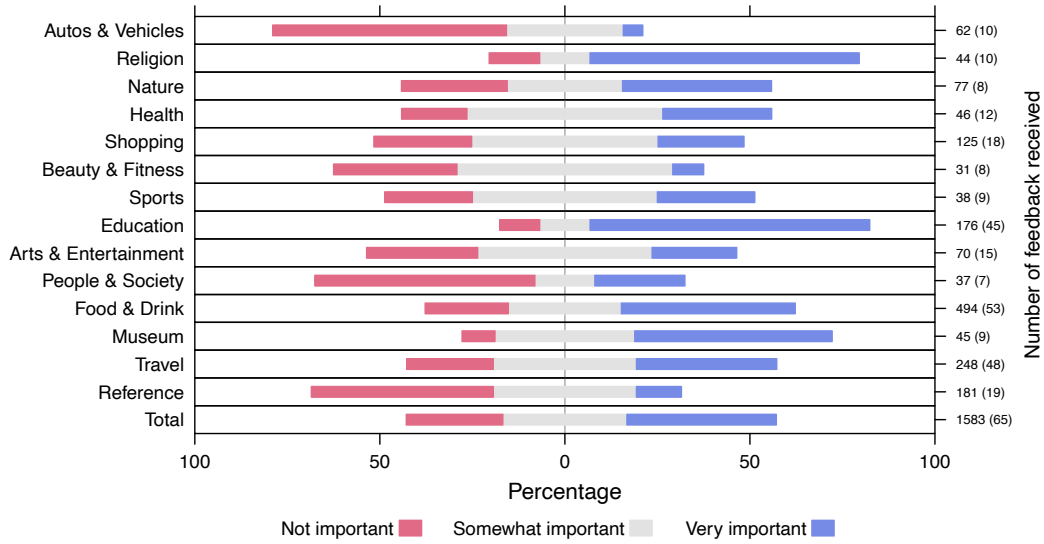

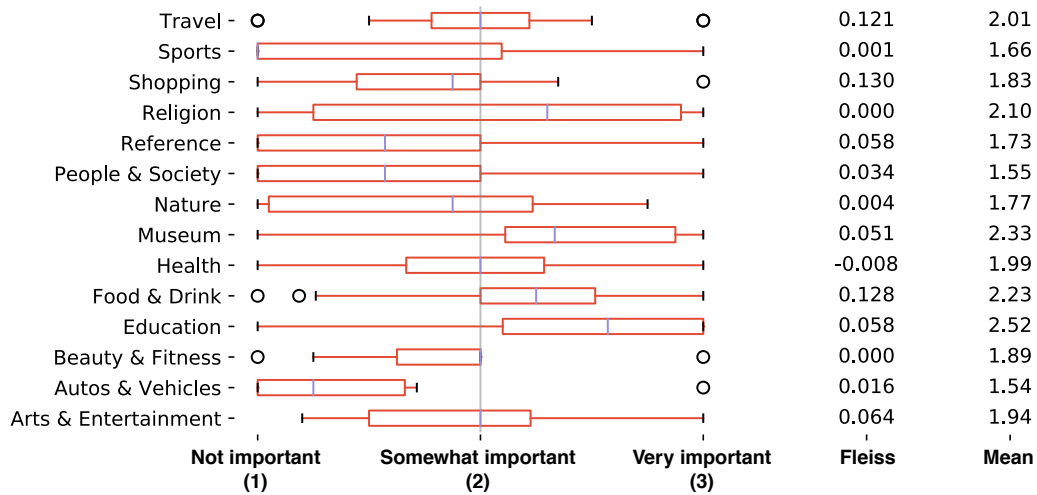Fig. 8. Perceived importance of topic of interest, classified by high-level categories.

Fig. 9. Box plot of the user feedback distribution and Kappa test results for the privacy sensitivity divergence per topic of interest categories with the corresponding Fleiss' Kappa and average mean of user feedback scores.

In Figure 8, we show a breakdown of the perceived importance of each topic of the different Topic of Interest. From this chart, Religion and Education are two topics that are considered important by the users, with respectively 72% and 76% of the feedback rating the topics as *Very important*. The fact that Religion is considered important matches with our previous observation of the perceived importance of related category Religious Belief in Figure 6. However, the other topics are considered as somewhat important by the users, which can denote a possible too broad categorization and labelling of the places the users visited *e.g.*, with the topic Arts & Entertainment, or that the users are finding the topic genuinely non important, *e.g.*, with Autos & Vehicles. This is further confirmed when inspecting the level of agreement of the users, reported in Figure 9. The topic Arts & Entertainment has no agreement ($\kappa = 0.064$) and a high amplitude in its perceived importance, which denotes diverging opinions. In the contrary, the topic Autos & Vehicles has low amplitude and is considered as not important by the users.
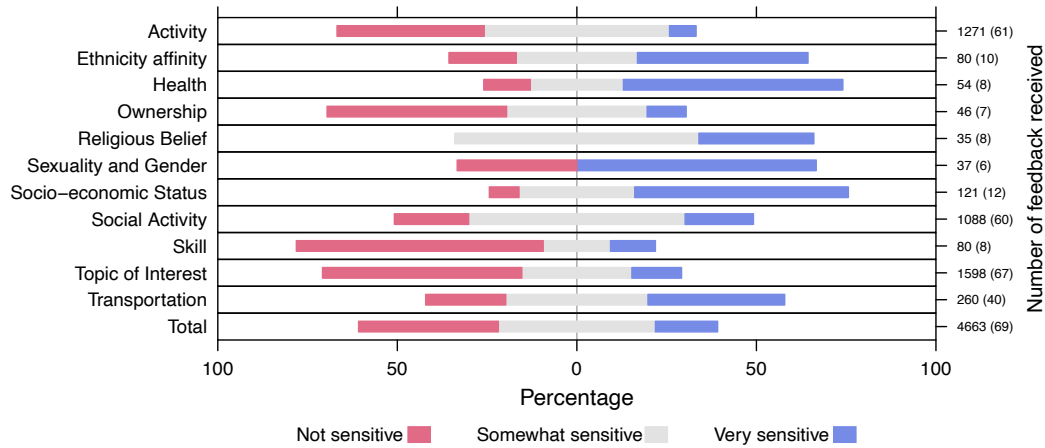
4.5 Personal Information Privacy Sensitivity

Fig. 10. Privacy sensitivity of aggregated personal information.

Perceived Privacy Sensitivity of Aggregated Personal Information Items. In Figure 10, we represent the feedback we collected from the users about their perception of the privacy sensitivity of the different personal information items. We also use a diverging stacked bar chart to represent the information. Overall, most of the personal information extracted from the places visited is considered as not sensitive or somewhat sensitive, with 36% and 47% of the collected feedback, respectively. This information includes user Activity (95%), Topic of interest (86%), Ownership (85%), Skill (81%), and Social activity (81%).

In fact, only 17% of the personal information is considered very sensitive by the users. The personal information categories that are considered most sensitive are Health (61%), Socio-economic Status (58%), Ethnic affinity (50%), Transportation (38%), and Religious Belief (32%). These categories of personal information are, intuitively, the most sensitive in terms of privacy.

Figure 11 provides a statistical analysis of the agreement among users of the perceived privacy sensitivity of the personal information items. From the boxplots, we see that the majority of the categories do not have a clear privacy sensitivity preference, except Topic of interest and Activity. Again, this is probably due to the fact that these categories are too broad and there is a high divergence among the privacy preferences of the subcategories. We will analyze the differences among the topics of interest next. The lack of clear preferences is also confirmed

Fig. 11. Box plot of the user feedback distribution and Kappa test results for the privacy sensitivity divergence per personal information categories with the corresponding Fleiss' kappa and average mean of user feedback.

by the Kappa test and no agreement, as κ is close to 0 for the majority of the categories. These results suggest that personalized settings associated to user profiles might be a suitable choice for capturing different privacy expectations of individuals.

Fig. 12. Privacy sensitivity of the topics of interest, classified by high-level categories.

Perceived Privacy Sensitivity of the Topics of Interest. Next, we inspect the topics of interest, classified by high-level categories represented with the diverging stacked bar chart in Figure 12. In this chart, we notice that,

	Not sensitive (1)	Somewhat sensitive (2)	Very sensitive (3)	Fleiss	Mean
Travel				0.037	1.40
Sports				0.015	1.42
Shopping				0.017	1.48
Religion				0.000	2.27
Reference				0.049	1.43
People & Society				0.010	1.31
Nature				0.012	1.48
Museum				0.009	1.42
Health				0.000	2.29
Food & Drink				0.079	1.29
Education				0.022	1.39
Beauty & Fitness				0.000	1.46
Autos & Vehicles				0.058	1.20
Arts & Entertainment				0.029	1.52

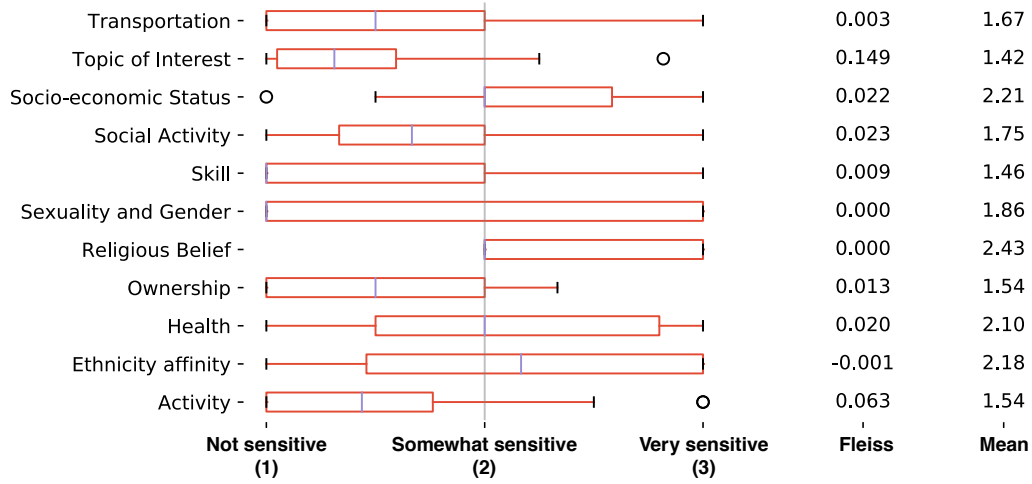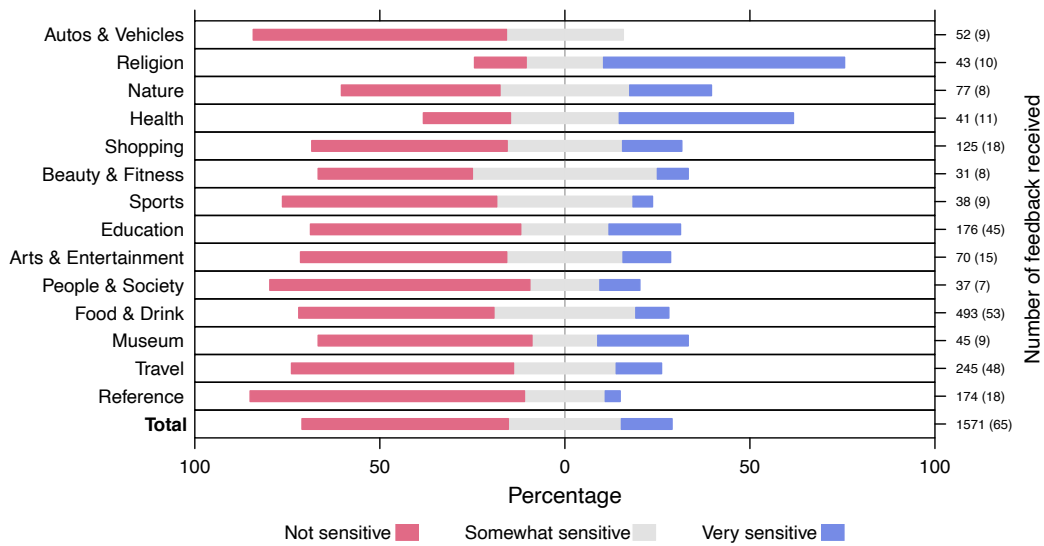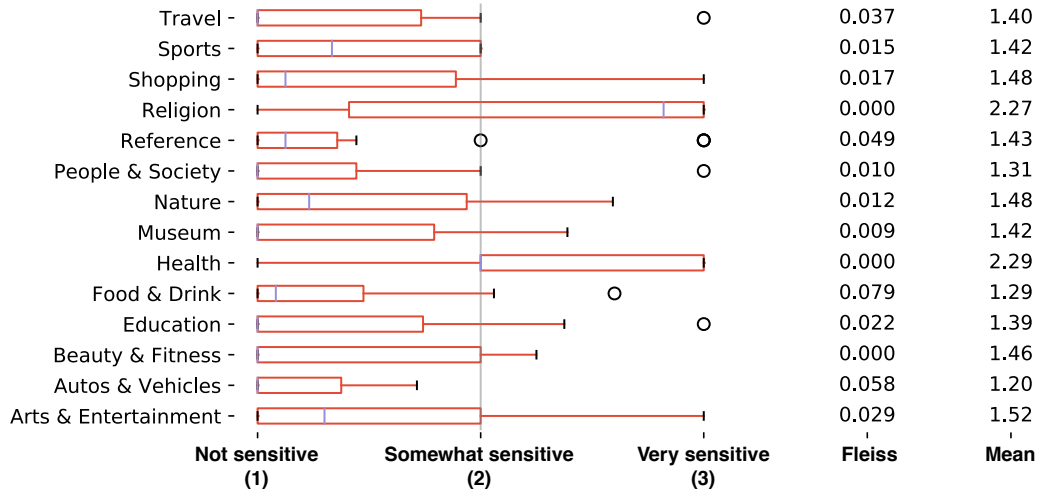Fig. 13. Box plot of the user feedback distribution and Kappa test results for the privacy sensitivity divergence per topic of interest categories with the corresponding Fleiss' kappa and average mean of user feedback.

while most of the interest categories are not considered as sensitive in terms of privacy by the users, the topics of Religion and Health are considered as very sensitive (72% and 50%, respectively).

Using Figure 13, we can see from the boxplots that, while most topic of interest are considered as no-to-somewhat sensitive by the users, Religion and Health are considered as sensitive, which is consistent with the privacy preferences observed in Figure 10 for the Religious Belief and Health categories. While there is either a general trend towards *Not sensitive* or *Very sensitive*, the Kappa test shows that there is no agreement among the users. This further confirms that privacy preferences are personal and no general trend can be derived.

4.6 Personal Information Importance and Privacy Sensitivity

We conclude our analysis by examining the closeness of the perceived importance and the privacy sensitivity of the personal information items. To this end, we gather the feedback for the aggregated personal information items for all users. We then measure the feedback correlation using the Pearson's correlation coefficient.

Table 3 confirms that the privacy sensitivity of a personal information item closely linked to its perceived importance by the users. This is even more apparent for categories such as Health and Religious Beliefs, which are considered as privacy sensitive.

From our analysis of the participants' feedback, we have surfaced that some personal information categories such as Health and Religious Beliefs are considered as sensitive by the majority of the users in terms of privacy. While other items are generally considered as less privacy sensitive, we noted that the privacy preferences vary among the users. Indeed, we have shown that this variation in the privacy sensitivity is highly related to the perceived importance for the personal information items. Finally, these findings confirm those illustrated in Goldbeck *et al.*'s study [29] and suggest that except for topics such as health and religion, we cannot build general privacy preferences, and we have to learn each user's personal privacy preferences over time.

Table 3. Correlation between the perceived importance and privacy expectation for personal information categories.

Personal information	Correlation
Activity	0.232
Ethnicity affinity	0.454
Health	0.781
Ownership	0.325
Religious Belief	0.796
Sexuality and Gender	0.729
Skill	0.300
Social Activity	0.431
Socio-economic Status	0.462
Topic of Interest	0.562
Transportation	0.343

5 DISCUSSION AND IMPLICATIONS

5.1 Limitations of Current Privacy-Preserving Systems

Privacy is a key concern for users, who generally express a preference for being in control of data sharing processes and related protection mechanisms [29]. For this reason, the focus of several projects has been on the problem of personal information inference from location information formulated as an *inference attack* [40, 41, 55]. This assumes that an attacker is able to access the location information of users, usually disclosed while accessing location-based services. A large number of solutions that aim at mitigating this type of attack have been proposed, such as anonymity and obfuscation techniques, as well as differential privacy methods [24, 32, 41, 55, 62]. For example, in [62] Yang *et al.* proposed PrivCheck, a privacy-preserving framework that also protects users from inference attacks by obfuscating their location data if given personal information (*e.g.*, the user's gender) can be leaked to third parties. This framework aims at minimizing the sensitive data that is leaked while maximizing the utility of the obfuscated data, that is the degree to which the user data can be used with a location-based service, for instance place recommendations that are personalized according to the user's preference.

These privacy-preserving techniques address the problem of inference attacks by proposing general-purpose frameworks that mitigate the privacy disclosure risk usually by adding an extra pre-processing layer. The challenge is to port these frameworks to the plethora of available devices that are available today, such as phones, tablets, gaming consoles, smartmeters, smartwatches, and so on. In fact, it is not possible to directly modify the application's source code and implement system-level libraries on non-rooted devices. End-users must therefore trust the willingness of application developers and operating systems to protect their privacy. As so, it is unlikely that applications, and in particular advertising and location libraries, will implement such framework that may prevent them from collecting high quality fine-grained data.

The privacy-preserving frameworks presented assumes that thee is a fixed set of personal information already known and specified. As we will see below, this study shows that it is important to learn privacy preferences that are specific to the users, i.e, without making an a priori assumption that users are only interested in protecting a given set of personal information. The frameworks automatically take actions to obfuscate or delete part of the user traces, without providing any recommendations or explanations as to how the traces that the user generated can violate the privacy expectations that they set. Finally, we showed that the use of a third party entity to protect the user's privacy can only be viable if the user trusts the privacy recommendations. It has been shown that explanations greatly improves the trust of the recommendations, and the willingness of users to make decisions based on them [42].

5.2 Design Guidelines for Future Privacy-Preserving Systems

This study provides insights to set design guidelines for future privacy-preserving mobile systems. It confirms that people perceive different kinds of personal information with varying degrees of privacy preferences, as discussed in [28]. Our work advances the state-of-the-art by showing that it is important to understand user privacy expectations and preferences with respect to the personal information that can be inferred from places. This understanding is key to design relevant recommendation systems that help users protect their privacy with respect to their own expectations at a fine-grain level. Automatic recommendations could warn the users that personal information items they find sensitive can be inferred in the proximity of places they visit. As a result, users who prefer that personal information related to health remain private could be interested in having recommendations to protect the disclosure of this type of personal information. An example could be a system that automatically notifies the user about potential disclosure risks when the user visits specific places, which might reveal sensitive personal information. For example, users could receive notifications whenever they are in the proximity of a doctor's clinic or a hospital. Alternatively, a collection of privacy-enhancing technologies could automatically obfuscate the data based on predefined privacy settings before sharing it with third parties. In our final survey, we showed that users value this automatic understanding of fine-grained personal information and they could be happy to take the time to give their feedback about their own expectations. This continuous feedback is key to understand individual privacy expectations and provide personalized recommendations in an online learning manner that improves with user feedback. In addition to recommendations that are personalized according to their privacy expectations and preferences, users highlighted the importance of explanations for the personal information inferences. When asked to give feedback on the explanations as it is shown in Figure 3d, only 6% of the users did not find the explanations useful. Instead 39% and 55% of users found them useful and very useful, respectively. Explanations are key to make the inferences interpretable and enable users to trust the recommendations to help the user protect their privacy. There is still limited work in this area. An initial proposal for an architecture for privacy-preserving pervasive systems has been recently presented in [7].

5.3 Methodological Considerations and Limitations

Throughout our study, we took several steps to ensure that users provided us with feedback valuable for our analysis. This has proven to be challenging, given the number of review tasks and the lack of financial incentives. At the same time, the lack of financial motivation might have avoided the symmetric problem of users providing information only for obtaining or increasing their compensation. Some participants may have purposefully given the wrong information or voluntarily not confirmed visits at places, for instance for privacy reasons. To this end, we designed our application to send daily notifications to remind the users to complete the different review tasks they have not completed yet. One design goal was to ensure that the users had all the necessary information to understand the goal and requirements of the study itself. The minimum two-week continuous participation filters out the participants who did not wish to participate in the study.

As presented and evaluated in the Appendix B, the inferences made by our backend system about the places visited by users and the personal information extracted from them was not entirely accurate. We provided UI interfaces to give the possibility to the users to correct the place and add or modify the corresponding personal information in order to collect ground truth data. However, the analysis of the feedback is based only on information that the users validated, including their visits at places and the personal information extracted from the places they visited.

The analysis we provided in our study is based on ground truth data that was validated by the participants directly from our application. We used this data to measure the importance and the privacy sensitivity of the personal information extracted from the places in the context of information sharing with unknown third parties such as advertisers. However, these third parties have rarely access to such ground truth data that has been

validated by the users and need to rely on the sole GPS location to infer the visits at places and provided tailored content. These entities can rely on similar or more sophisticated place extraction algorithm than the one we used in our study to infer the places, but they can also exploit the fact that users return to places that interest them, including home, work, and other third places such as supermarkets or coffee shops. The accuracy of the place inference can then improve with the repetition of visits despite the lack of user feedback.

Finally, we argue that the methodology of this study can be generalized to a study with an extended user base, in terms of number of participants but also not limited to iPhone users, but also Android users, who might have different privacy preferences and other visit patterns. Indeed, different populations of users might provide additional insights about places that have fewer visits in the dataset used for this study, *e.g.*, health-related places.

In Section 4 we further observed low Kappa values, which denote no agreement in the majority of the cases with the perceived importance and privacy sensitivity of the personal information categories. While this large response variability indicates a need to have fine-grained personalized privacy preferences, it also suggests that the personal information categories were too broad, especially when no clear agreement was reached among the participants.

6 CONCLUSION AND FUTURE WORK

Users routinely share their location with a wide range of services and advertisers through mobile applications and libraries installed on their phones. We have conducted a novel *in-the-wild* research study to understand the range of personal information that can be inferred from continuous location tracking, and whether these are perceived as privacy-sensitive by the participants. To this end, we have developed TrackingAdvisor, a mobile application that collects the location of the participants in the background, automatically extracts the places they visited, and infers the related personal information. Through the app, users can also provide their feedback on the relevance, importance, and privacy sensitivity of the personal information extracted from location traces. We have showed that while some personal information categories such as the user's activities, skills, and social activities are not perceived as sensitive in terms of privacy, while other categories of personal information such as health, religious belief, ethnic affinity, and socio-economic status are considered private by the participants of the study.

We would also like to point out that one of the key contributions of this work is methodological. We believe that the methodology used in this study can be adopted for future studies with different populations and contexts. This study paves the way to the design and implementation of the next generation of privacy-preserving ubiquitous systems. We believe that understanding fine-grain privacy sensitivities of personal information is key to making relevant recommendations that help users take useful actions to protect themselves against unwanted personal information inferences. In particular, we have showed that explanations are important for users to understand how and why specific personal information can be inferred [23]. For this reason, our findings suggest that recommendations should include explanations to raise user awareness of the range of personal information that can be inferred with their data and to enable them to make informed decisions as to whether to share their data or not. We believe that interpretability should be one of the key drivers of the design of future privacy-preserving systems.

ACKNOWLEDGMENTS

REFERENCES

[1] Foursquare. https://foursquare.com.

[2] Yuvraj Agarwal and Malcolm Hall. ProtectMyPrivacy: Detecting and Mitigating Privacy Leaks on iOS Devices using Crowdsourcing. In *ACM MobiSys*, pages 97–110, Taipei, Taiwan, June 2013. ACM.

[3] Laura Alessandretti, Piotr Sapiezynski, Vedran Sekara, Sune Lehmann, and Andrea Baronchelli. Evidence for a Conserved Quantity in Human Mobility. *Nature Human Behaviour*, 2(7):485–491, 2018.

[4] Hazim Almuhimedi, Florian Schaub, Norman Sadeh, Idris Adjerid, Alessandro Acquisti, Joshua Gluck, Lorrie Faith Cranor, and Yuvraj Agarwal. Your Location Has Been Shared 5,398 Times!: A Field Study on Mobile App Privacy Nudging. In *ACM CHI*, pages 787–796, Seoul, Republic of Korea, April 2015. ACM.

[5] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. DBPedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer, 2007.

[6] Louise Barkhuus and Anind K Dey. Location-Based Services for Mobile Telephony: A Study of Users' Privacy Concerns. In *INTERACT*, volume 3, pages 702–712, Zurich, Switzerland, 2003. IFIP.

[7] Benjamin Baron and Mirco Musolesi. Interpretable Machine Learning for Privacy-Preserving Pervasive Systems. *IEEE Pervasive Computing*, 19(01):73–82, January 2020.

[8] Bin Bi, Milad Shokouhi, Michal Kosinski, and Thore Graepel. Inferring the Demographics of Search Users: Social Data Meets Search Queries. In *WWW*, pages 131–140, Rio de Janeiro, Brazil, May 2013. ACM.

[9] Steven Bird and Edward Loper. NLTK: The Natural Language toolkit. In *ACL*, page 31, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[10] AJ Brush, John Krumm, and James Scott. Exploring End-User Preferences for Location Obfuscation, Location-Based Services, and the Value of Location. In *ACM UbiComp*, pages 95–104, Copenhagen, Denmark, September 2010. ACM.

[11] Luca Canzian and Mirco Musolesi. Trajectories of Depression: Unobtrusive Monitoring of Depressive States by Means of Smartphone Mobility Traces Analysis. In *ACM UbiComp*, pages 1293–1304, Osaka, Japan, September 2015. ACM.

[12] Saksham Chitkara, Nishad Gothoskar, Suhas Harish, Jason I Hong, and Yuvraj Agarwal. Does this App Really Need My Location?: Context-Aware Privacy Management for Smartphones. *ACM IMWUT*, 1(3):42, 2017.

[13] Martin J Chorley, Roger M Whitaker, and Stuart M Allen. Personality and Location-Based Social Networks. *Computers in Human Behavior*, 46:45–56, 2015.

[14] Sunny Consolvo, Ian E Smith, Tara Matthews, Anthony LaMarca, Jason Tabert, and Pauline Powledge. Location Disclosure to Social Relations: Why, When, & What People Want to Share. In *ACM CHI*, pages 81–90, Portland, OR, USA, April 2005. ACM.

[15] Mihaly Csikszentmihalyi and Reed Larson. Validity and Reliability of the Experience-sampling Method. In *Flow and the foundations of positive psychology*, pages 35–54. Springer, 2014.

[16] Aron Culotta, Nirmal Kumar Ravi, and Jennifer Cutler. Predicting the Demographics of Twitter Users from Website Traffic Data. In *AAAI*, pages 72–78, Austin, TX, USA, January 2015. AAAI Press.

[17] Dan Cvrcek, Marek Kumpost, Vashek Matyas, and George Danezis. A Study on the Value of Location Privacy. In *ACM WPES*, pages 109–118, Alexandria, VA, USA, October 2006. ACM.

[18] George Danezis, Stephen Lewis, and Ross J Anderson. How Much is Location Privacy Worth? In *WEIS*, volume 5, 2005.

[19] Yves-Alexandre De Montjoye, César A Hidalgo, Michel Verleysen, and Vincent D Blondel. Unique in the Crowd: The Privacy Bounds of Human Mobility. *Scientific reports*, 3:1376, 2013.

[20] Riccardo Di Clemente, Miguel Luengo-Oroz, Matias Travizano, Sharon Xu, Bapu Vaitla, and Marta C González. Sequences of Purchases in Credit Card Data Reveal Life Styles in Urban Populations. *Nature communications*, 9, 2018.

[21] EU Directive. 95/46/ec of the european parliament and of the council of 24 october 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. *Official Journal of the EC*, 23(6), 1995.

[22] EU Directive. Article 4 (2). *General Data Protection Regulation (GDPR)*, 2018.

[23] Finale Doshi-Velez and Been Kim. Towards a Rigorous Science of Interpretable Machine Learning. *arXiv preprint arXiv:1702.08608*, 2017.

[24] Cynthia Dwork. Differential Privacy: A Survey of Results. In *International Conference on Theory and Applications of Models of Computation*, pages 1–19. Springer, 2008.

[25] Adrienne Porter Felt, Elizabeth Ha, Serge Egelman, Ariel Haney, Erika Chin, and David Wagner. Android Permissions: User Attention, Comprehension, and Behavior. In *ACM SOUPS*, page 14, Washington, DC, USA, July 2012. ACM.

[26] Joseph L Fleiss. Measuring Nominal Scale Agreement Among Many Raters. *Psychological bulletin*, 76(5):378, 1971.

[27] Sébastien Gambs, Marc-Olivier Killijian, and Miguel Núñez del Prado Cortez. Show Me How You Move and I Will Tell You Who You Are. In *ACM SIGSPATIAL International Workshop on Security and Privacy in GIS and LBS*, pages 34–41, San Jose, CA, USA, November 2010. ACM.

[28] Erving Goffman et al. *The Presentation of Self in Everyday Life.* Harmondsworth London, 1959.

[29] Jennifer Golbeck and Matthew Louis Mauriello. User Perception of Facebook App Data Access: A Comparison of Methods and Privacy Concerns. *Future Internet*, 8(2):9, 2016.

[30] Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. Understanding Individual Human Mobility Patterns. *nature*, 453(7196):779, 2008.

[31] Thore Graepel, Joaquin Quinonero Candela, Thomas Borchert, and Ralf Herbrich. Web-Scale Bayesian Click-Through Rate Prediction for Sponsored Search Advertising in Microsoft's Bing Search Engine. In *ICML*, Haifa, Israel, June 2010.

[32] Marco Gruteser and Xuan Liu. Protecting Privacy, in Continuous Location-Tracking Applications. *IEEE Security & Privacy*, 2(2):28–34, 2004.

[33] Ramanathan Guha, Vineet Gupta, Vivek Raghunathan, and Ramakrishnan Srikant. User Modeling for a Personal Assistant. In *ACM WSDM*, pages 275–284, Shanghai, China, February 2015. ACM.

[34] Mordechai Haklay and Patrick Weber. OpenStreetMap: User-generated Street Maps. *IEEE Pervasive Computing*, 7(4):12–18, 2008.

[35] Ramaswamy Hariharan and Kentaro Toyama. Project Lachesis: Parsing and Modeling Location Histories. In *GIScience*, pages 106–124, Adelphi, MD, USA, October 2004. Springer.

[36] John A Hartigan and Manchek A Wong. Algorithm AS 136: A k-means Clustering Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108, 1979.

[37] Iris A Junglas and Richard T Watson. Location-Based Services. *Communications of the ACM*, 51(3):65–69, 2008.

[38] Ashraf Khalil and Kay Connelly. Context-aware Telephony: Privacy Preferences and Sharing Patterns. In *ACM CSCW*, pages 469–478, Banff, Alberta, Canada, November 2006. ACM.

[39] Michal Kosinski, David Stillwell, and Thore Graepel. Private Traits and Attributes Are Predictable from Digital Records of Human Behavior. *Proceedings of the National Academy of Sciences*, 110(15):5802–5805, 2013.

[40] John Krumm. Inference Attacks on Location Tracks. In *International Conference on Pervasive Computing*, pages 127–143, White Plains, NY, USA, July 2007. Springer.

[41] John Krumm. A Survey of Computational Location Privacy. *Personal and Ubiquitous Computing*, 13(6):391–399, 2009.

[42] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. Principles of Explanatory Debugging to Personalize Interactive Machine Learning. In *ACM IUI*, pages 126–137, Atlanta, GA, USA, 2015. ACM.

[43] Quannan Li, Yu Zheng, Xing Xie, Yukun Chen, Wenyu Liu, and Wei-Ying Ma. Mining User Similarity Based on Location History. In *ACM SIGSPATIAL*, page 34, Irvine, CA, USA, November 2008. ACM.

[44] Ilaria Liccardi, Alfie Abdul-Rahman, and Min Chen. I Know Where You Live: Inferring Details of People's Lives by Visualizing Publicly Shared Location Data. In *ACM CHI*, pages 1–12, San Jose, CA, USA, May 2016. ACM.

[45] Erika McCallister. *Guide to Protecting the Confidentiality of Personally Identifiable Information.* Diane Publishing, 2010.

[46] Mary L McHugh. Interrater Reliability: The Kappa Statistic. *Biochemia medica*, 22(3):276–282, 2012.

[47] K Mohamed, Etienne Côme, Johanna Baro, and Latifa Oukhellou. Understanding Passenger Patterns in Public Transit Through Smart Card and Socioeconomic Data. In *UrbComp*, Seattle, WA, USA, September 2014. ACM.

[48] David C Mohr, Mi Zhang, and Stephen M Schueller. Personal Sensing: Understanding Mental Health using Ubiquitous Sensors and Machine Learning. *Annual Review of Clinical Psychology*, 13:23–47, 2017.

[49] Feng Qiu and Junghoo Cho. Automatic Identification of User Interest for Personalized Search. In *WWW*, pages 727–736, Edinburgh, UK, May 2006. ACM.

[50] Suzanne Rivoire, Mehul A Shah, Parthasarathy Ranganathan, and Christos Kozyrakis. JouleSort: A Balanced Energy-Efficiency Benchmark. In *ACM SIGMOD*, pages 365–376, Beijing, China, June 2007.

[51] Naomi B Robbins, Richard M Heiberger, et al. Plotting Likert and other Rating Scales. In *Joint Statistical Meeting*, pages 1058–1066, Alanta, GA, USA, July 2001. American Statistical Association.

[52] Luca Rossi and Mirco Musolesi. It's the Way You Check-In: Identifying Users in Location-Based Social Networks. In *ACM COSN*, Dublin, Ireland, October 2014. ACM.

[53] Fuming Shih, Ilaria Liccardi, and Daniel Weitzner. Privacy Tipping Points in Smartphones Privacy Preferences. In *ACM CHI*, pages 807–816, Seoul, Republic of Korea, April 2015. ACM.

[54] Milad Shokouhi. Learning to Personalize Query Auto-Completion. In *ACM SIGIR*, pages 103–112, Dublin, Ireland, July 2013. ACM.

[55] Reza Shokri, George Theodorakopoulos, Jean-Yves Le Boudec, and Jean-Pierre Hubaux. Quantifying Location Privacy. In *IEEE Symposium on Security and Privacy*, pages 247–262, Oakland, CA, USA, May 2011. IEEE.

[56] Einar Snekkenes. Concepts for Personal Location Privacy Policies. In *ACM EC*, pages 48–57, Tampa, Florida, USA, October 2001. ACM.

[57] Jacopo Staiano, Nuria Oliver, Bruno Lepri, Rodrigo de Oliveira, Michele Caraviello, and Nicu Sebe. Money walks: A Human-centric Study on the Economics of Personal Mobile Data. In *ACM UbiComp*, pages 583–594, Seattle, WA, USA, 2014. ACM.

[58] Jaime Teevan, Susan T Dumais, and Eric Horvitz. Personalizing Search via Automated Analysis of Interests and Activities. In *ACM SIGIR*, pages 449–456, Salvador, Brazil, August 2005. ACM.

[59] Eran Toch, Justin Cranshaw, Paul Hankes Drielsma, Janice Y Tsai, Patrick Gage Kelley, James Springfield, Lorrie Cranor, Jason Hong, and Norman Sadeh. Empirical Models of Privacy in Location Sharing. In *ACM UbiComp*, pages 129–138, Copenhagen, Denmark, September 2010. ACM.

[60] Janice Y Tsai, Patrick Kelley, Paul Drielsma, Lorrie Faith Cranor, Jason Hong, and Norman Sadeh. Who's Viewed You?: The Impact of Feedback in a Mobile Location-Sharing Application. In *ACM CHI*, pages 2003–2012, Boston, Massachusetts, USA, April 2009. ACM.

[61] Pengfei Wang, Jiafeng Guo, Yanyan Lan, Jun Xu, and Xueqi Cheng. Your Cart Tells You: Inferring Demographic Attributes from Purchase Data. In *ACM WSDM*, pages 173–182, San Francisco, CA, USA, February 2016. ACM.

[62] Dingqi Yang, Daqing Zhang, Bingqing Qu, and Philippe Cudré-Mauroux. PrivCheck: Privacy-Preserving Check-In Data Publishing for Personalized Location-Based Services. In *ACM UbiComp*, pages 545–556, Heidelberg, Germany, 2016. ACM.

[63] Paul A Zandbergen. Accuracy of iPhone Locations: A comparison of Assisted GPS, WiFi and Cellular Positioning. *Transactions in GIS*, 13(1):5–25, 2009.

[64] Lei Zhang, Jiangchuan Liu, Hongbo Jiang, and Yong Guan. SensTrack: Energy-Efficient Location Tracking with Smartphone Sensors. *IEEE Sensors Journal*, 13(10):3775–3784, 2013.

[65] Yu Zheng, Xing Xie, and Wei-Ying Ma. Geolife: A Collaborative Social Networking Service Among User, Location and Trajectory. *IEEE Data(base) Engineering Bulletin*, June 2010.

[66] Erheng Zhong, Ben Tan, Kaixiang Mo, and Qiang Yang. User Demographics Prediction based on Mobile Data. *Pervasive and Mobile Computing*, 9(6):823–837, 2013.

[67] Yuan Zhong, Nicholas Jing Yuan, Wen Zhong, Fuzheng Zhang, and Xing Xie. You Are Where You Go: Inferring Demographic Attributes from Location Check-Ins. In *ACM WSDM*, pages 295–304, Shanghai, China, February 2015. ACM.

[68] Xiaoyong Zhou, Soteris Demetriou, Dongjing He, Muhammad Naveed, Xiaorui Pan, XiaoFeng Wang, Carl A Gunter, and Klara Nahrstedt. Identity, Location, Disease and More: Inferring your Secrets from Android Public Resources. In *ACM CCS*, pages 1017–1028, Berlin, Germany, November 2013. ACM.

A USER LOCATION COLLECTION

In order to detect the places that the users are visiting with high precision, we need to design a system that collects locations with a high accuracy. However, achieving an accuracy of a few meters requires the use of the GPS receiver, which has a high power consumption, as compared to other modalities such as Bluetooth, WiFi, and cellular that are less power-hungry but also less accurate [64]. Since energy-efficiency has become a crucial requirement in terms of application usability [50], we design our location collection system such that it trades energy-efficiency with the level of accuracy required for the specific needs of our application: (i) Continuous tracking of locations in the background at regular intervals, even after periods of inactivity, (ii) high accuracy detection of visits at places, and (iii) automatic pausing of the continuous tracking when the user has stopped moving (*i.e.*, the user is visiting a place).

The Core Location framework provided by iOS offers different location modes to retrieve user locations, with various levels of accuracy and energy efficiency, namely (1) Continuous updates retrieves locations at a high frequency and accuracy, (2) Significant updates retrieves locations when the user position changes by a significant amount (*e.g.*, 500 meters), (3) Region updates monitors entrances and exits at configurable geofences, and (4) Visit updates detects visits at locations where the user has stationed for some time (typically 5 minutes). While the Continuous updates mode has the most energy impact, it also gives the most accurate user location. The other modes have both low accuracy of a few hundreds of meters and low energy impact. Since none of the modes can satisfy all of the requirements listed above, our system consists of a hybrid solution that leverages the different location modes.

One of the main challenges we had to face was to keep the user location collection system running in the background. Indeed, iOS constrains the duration of the background state of applications and can suspend them in order to preserve battery life, killing any timers or long-running tasks that were active. However, the Significant updates, Region updates, and Visit updates modes all overcome this limitation by triggering an application launch in the background to process the retrieved locations. We track the user location at regular intervals by using the Region updates mode by creating a region of a given size (100 meters) centered at the user's current location. When the user has exited the region, the system updates the center of the region to the new user location retrieved using the Continuous updates mode for a short period of time. Note that this period of time

should be long enough to obtain an accurate location and short enough to have a low energy impact. Empirically, we found that allowing the Continuous updates mode to run for one second gives a good trade-off.

With this setting, we need to retrieve an accurate user location within the current region in order to match the place visited by the user. To this end, we set a two-minute timer before requesting a new user location using Continuous updates mode. This gives sufficient time for the user to either exit the region at a normal pace (*i.e.*, 1.4 m/s) or visit a place within the region. We also need to account for the variability of the user's speed, depending on the different modes of transportation they may use. Indeed, at higher speeds users will pass through several regions, triggering frequent Continuous updates, which has a significant energy impact. To mitigate this problem, the system adaptively determines the radius of the current region depending on the user's displacement speed (calculated using the user's previous recorded location) such that the location update frequency remains the same.

Due to the fact that the Region updates mode relies on the cellular network to detect region exits, it might lack precision as users may exit a region without triggering any updates (*e.g.*, if the user is traveling at high speed). To this end, we leverage both the Significant updates and Visit updates modes to trigger app launches and create a new region around the user's new location. The application periodically sends user locations to the back-end server in the background, when connected to WiFi networks in order to save cellular data.

B PLACE EXTRACTION

In this Appendix, we provide some key technical details about place extraction, which are not central to the goal of this work. However, we believe that it is important to provide this information in order to ensure full reproducibility of this study.

From the "raw" user locations collected by the TrackingAdvisor application, the goal of the place extraction is (1) to detect the stay points, which correspond to stay points where users dwelled for a minimum amount of time, and (2) to match them to actual places using open-access place databases. Without loss of generality, we used OpenStreetMap [34] and Foursquare [1], but our place matching algorithm can leverage any open-access place database. In the following, we detail the stay point detection and place matching algorithms.

B.1 Stay Point Detection

The development and general availability of GPS trackers has opened new avenues for research with location traces, in particular to study and understand the context of these locations and their significance to the users.

Ashbrook and Starner proposed one of the first algorithms to automatically detect clusters from raw location points. They proposed this algorithm as part of a location model that aims to characterize the current user context (*e.g.*, user is at home) and create a probabilistic model of personal movement (*e.g.*, *what is the most likely place the user will go next?*). The algorithm aims to detect the places that have a significance to the user (*e.g.*, home or work), that is places where the user spends at least a certain amount of time from 1-5 minutes. The algorithm is a variant of the k-means clustering algorithm which finds the center of a place visited by the user by converging to the mean of the location points recorded at that place [36]. Hariharan and Toyama developed a different algorithm to detect stay points from traces of raw locations [35]. In particular, their algorithm takes into account the temporality of the traces to characterize stay points by a roaming distance and a stay duration. The algorithm extracts a stay point that corresponds to the medoid of the location points within a region of a given maximum radius if the stay duration within the region is greater than a given threshold. Typical stay duration thresholds are in the order of the minute, *e.g.*, a 5-minute stay duration within a region with a 100-meter radius corresponds to an actual visit at a place. The GeoLife dataset has paved the avenue to the design and development of a vast literature on extracting stay points from GPS traces [65]. Li *et al.* proposed an efficient

stay point detection algorithm similar to the algorithm developed by Hariharan and Toyama [43]. Instead of considering the medoid as the stay point, their algorithm computes the mean point as the stay point.

These algorithms have been developed with dense representations of user locations of typically one location point collected every 5-10 seconds and thus lead to good overall results when determining user stay points. However, in our case, we have a much more sparse representation of the user locations, which make the algorithms ill-suited to extract stay points from our traces. As so, with a minimum stay duration of 3 minutes and a maximum roaming distance of 100 meters, we had a lot of false positive results where stay points were detected at locations where the user was just passing by. We deliberately chose very conservative thresholds for the minimum stay duration and the maximum roaming distance in order to capture most of the places visited by the users and avoid missing places with short visits.

We propose modifications to Hariharan and Toyama's algorithm that mitigate the number of false positives and filter the stay points that were detected and did not match any places that the user have actually visited. In our real-world deployment, we encountered two distinct cases that lead to false positives. The first case happens when users slowly walk their way through an area, exceeding the stay duration threshold. In this case, the algorithm will detect a stay point, whereas the user did not stop at an actual place. To mitigate this case, we propose to modify the original algorithm by introducing an *inter-distance metric*. This metric allows us to separate the cases where users remain at the same place for a certain amount of time and the cases where the users slowly go through an area while exceeding the stay duration threshold. We compute the metric by averaging the distances between each successive point that have been recorded between the stay duration in the extracted stay point. If the average distance is greater than the roaming distance, we discard the stay point as a false positive. The second case happened when a user travels at speeds higher than the region update frequency, which creates a large number of points within the roaming distance that the algorithm detects as a stay point. To mitigate this case, we propose to modify the original algorithm by introducing an *average displacement speed metric*. We compute this metric by averaging the displacement speeds between each successive point recorded over the stay point duration. If the average displacement speed is greater than 1 m/s, that is just below the average walking speed, we discard the stay point as a false positive. With these two adaptations, we filter out the false positives that create non-relevant stay points.

B.2 Place Matching Algorithm

Once we have determined the stay points visited by the user, we need to match them to actual places. Our goal here is to obtain the most likely place that the user has visited based on the recorded location points associated to the stay point. To this end, we leverage place databases that contain various information about the place in a geographic environment. We use Foursquare and OpenStreetMap place databases to match the detected user stay points with the places they visited. In the following, we detail how we use both place databases. Note that, while we describe our place matching using the Foursquare database, we stress that the methodology can be applied to any place database.

When a user has just installed the application, we start by matching the stay points we detect from their location trace to the places contained in the Foursquare database. When the geometry information is available, the place from Foursquare are augmented by OpenSteetMap polygons to increase the accuracy of the place matching. In particular, if the polygon information is not available for the Foursquare place, the place matching algorithm falls back to a match based on the place's coordinates. Once a user has visited some places, we further use them when matching a stay point to a close-by previously visited place. In the following, we describe how we use the place databases for the place matching of a detected stay point.

We leverage Foursquare [1], a location-based social network, to match the detected stay points as we described in the previous section to actual places that the user has visited. Foursquare offers various APIs (Application

Programming Interfaces) to extract places — or venues — in the vicinity of a given location. Note that the place matching algorithm described in the following can be extended to include other place databases.

Location readings can be inaccurate, in particular in dense areas such as cities, airports, and malls, mostly because of the error in the location collected from the GPS, WiFi, and cellular signals. In addition to the different places and their geo-coordinates, Foursquare has valuable metadata about the places that we use to enhance our place matching algorithm, such as the number of check-ins, the average rating of the place, the number of likes, and the tips (a tip is a small review of a place given by the Foursquare users) associated to each place. While Foursquare offers an API to check-in at the most likely place given a location and the time of the request, this presents some disadvantages, as the request to retrieve the most likely place is made offline server-side, which can be well after the user has visited the place. Instead, we query all the places in the vicinity of the stay point visited by the user (typically, a 150-meter square centered in the stay point) and rank the places based on their distance from the center of the stay point and their popularity given by the number of check-ins and the time of the visit. Top-ranking places are (1) close to the center of the stay point, even if they do not have a large number of check-ins (the user may have visited a place that is not popular), and (2) very popular at the time of the visit, even though they are not too close to the center of the stay point.

B.3 Place Matching Algorithm Evaluation

We evaluate the performance of the place matching algorithm we explained in more detail in Appendix B.2 by examining the feedback we received from the users confirming the correctness of the place that was automatically extracted by our place matching algorithm. We rely on the feedback that the participants gave when confirming, deleting or correcting a visit directly from the timeline, as shown in Figure 2a with the button *Yes*, *Delete*, and *Correct*, respectively. In Figure 14a, we show the feedback we have received from the users in three distinct categories: (1) "Correct" represents the average proportion of times a user tapped on the button *Yes*, (2) "Wrong place" represents the average proportion of times a user tapped on the button *Correct*, and (3) "Not a visit" represents the average proportion of times a user tapped on the button *Delete*. As we can see on the graph, the users found that the visits that were automatically extracted were correct 57% of the time. False negatives (*i.e.*, "Not a visit") account for 27% of the visits extracted and the wrong places account for 16% of the visits. This shows the good overall performance of the place matching algorithm that gives a high place matching accuracy of 73.3% among the places with confirmed visits.

In order to understand the error made by the place matching algorithm, we represent in Figure 14b the cumulative distribution function of the distance between the incorrect location and that corrected by the user with the place search interface provided in the application (see Figures 2c and 2d). The median error distance is 94 meters, which corresponds to the typical error of the GPS readings in urban environments [63]. Also, since we relied on the geo-coordinates given by Foursquare, they may not correspond to the points were automatically collected by the application, as Foursquare uses the locations of the user check-ins to determine the coordinate of a place. These location may not be accurate in our setting, in particular for large venues not matched to any OpenStreetMap polygon.

B.4 Personal information inference

The goal of the personal information inference component is to extract and infer personal information items from the places users visited. To this end, we leverage place description and metadata from different place databases as detailed in the following.

While there are no established classifications of the different personal information, we structure them in broad categories such that they cover a large spectrum. The different personal information categories are listed

(a) Proportion of feedback received for visit confirmation.

(b) Cumulative distribution function of the distance of the place corrected by the user with respect to the place automatically inferred.
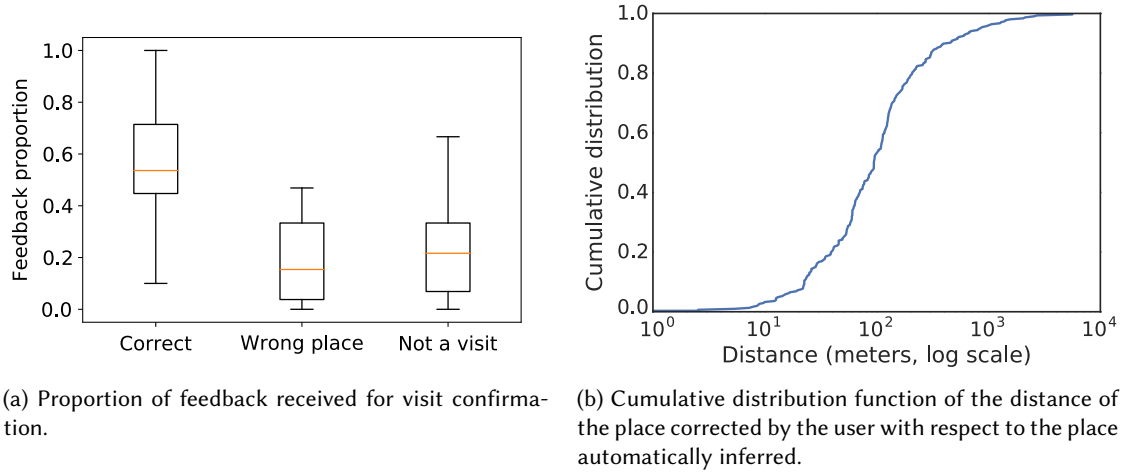
Fig. 14. Evaluation of the performance of the place matching algorithm.

in Table 4. We structure the personal information in broad categories such that they cover the large spectrum of personal information.

Socio-Economic and Demographic Information Inference. Information extracted from the Foursquare platform was used to infer the socio-economic status of the participants based on the places they visit. In particular, we propose to use the price tier of places that the participants visit to determine their socio-economic status, whether it is *low* or *high*. We specifically target retail places of the high-level categories Arts & Entertainment (*e.g.*, theaters), Food (*e.g.*, restaurants), Nightlife Spot (*e.g.*, bars), and Shop & Service (*e.g.*, department stores), as these places sell goods or services at different price ranges. The price tiers given by Foursquare associated to the place range from 1 (least pricey) to 4 (most pricey). As such, the idea is that if users visit a majority of places with a low price tier, they are more likely to have a low socio-economic status, whereas users who visit a majority of places with a high price tier are more likely to have a high socio-economic status. To this end, we label places with *low* socio-economic status if their price tier is 1 and with *high* socio-economic status if their price tier is 4.

We leverage the information provided by Foursquare along with the tips to infer the gender of the participants based on the places they visit. We compute the most likely gender (*i.e.*, male or female)[5] for each place category based on the proportion of check-ins generated by the male users and the female users. As ground truth, we consider the dataset collected by Yang *et al.* of 18,201 users in New York City during a period of about 18 months from April 2012 to September 2013 [62]. We plot the odds ratio for the two genders male and female from the users who have checked-in at the venues of the dataset in Figure 15. We consider that the place categories with an odds ratio greater than 2 for a given gender will be most likely visited by users of this gender. As such, we associate the gender information to the corresponding places. In Figure 15, we list the place categories with the greatest odds ratios for the female users (on the left-hand side) and for the male users (on the right-hand side).

Topics of Interest. We extract the topics of interest from the different data sources we considered. Since Foursquare has an extensive fine-grained place categorization of more than 900 place categories[6], we consider

---

[5]We are aware that this is a potential limitation of our proof of concept implementation. We used binary classification only because this is the information available in Foursquare, where it is possible only to select these two genders. However, please note that the algorithms used for the detection of male/female attributes can be generalized beyond this binary classification.

[6]Foursquare place categories. https://developer.foursquare.com/docs/resources/categories.
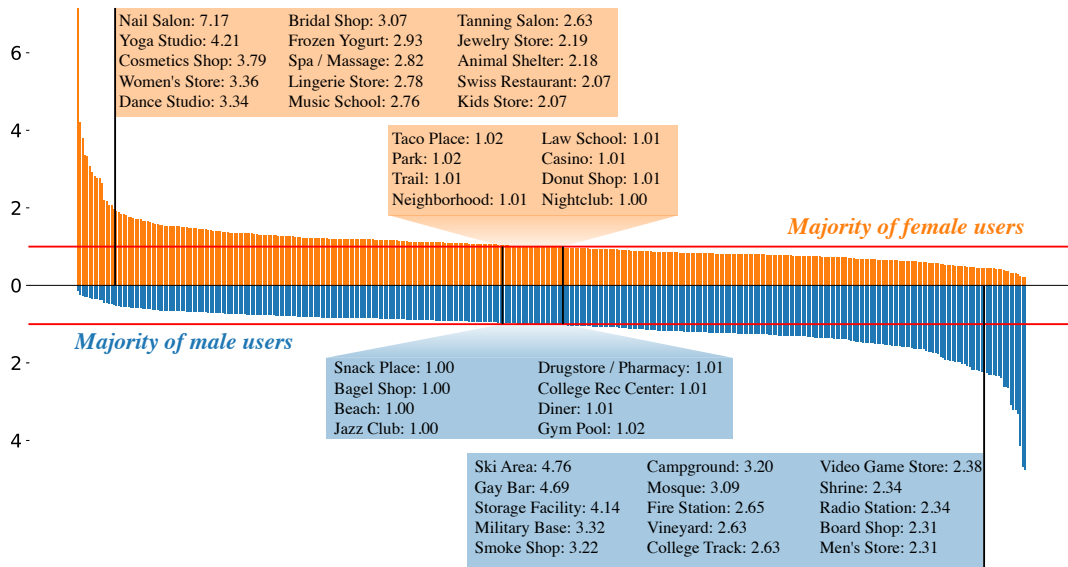
Fig. 15. Odds ratio for the different categories of places where Foursquare users of known genders have checked-in at least 100 times per category.

the place categories themselves as potential topics of interest to the users. Additionally, we also extract the information about the places contained in DBPedia as personal information [5]. DBPedia places are characterized by a set of types and related subjects. For instance, the Louvre museum has the types "Art museum" and "Historic site", as well as the related subjects, including "Archaeological museums in France", "History museums in France", and "National museums of France". These subjects characterize niche personal information about the user, such as their interests for archaeology or history. We process the different subjects per category to extract meaningful personal information using natural language processing (NLP) techniques. Specifically, for each place type (*e.g.*, museums, universities), we compute the most frequent root words extracted from the related subjects of each place of the given type with a dependency parser. For instance, the most frequent root words associated to the type "museum" are "gallery" and "institute". Through tokenization and lemmatization, we extract the most frequent *n*-grams from the related subjects and filter the *n*-grams that do not contain any of the extracted root words [9]. We further filter the *n*-grams ending or beginning with non-relevant adjectives or prepositions with a maximum likelihood estimator of a word sequence probability. As a result, *n*-grams such as "museum in new" or "museum in the united" will be filtered. We store the extracted *n*-grams as personal information and link them to the related place subjects from which they have been extracted. As a result, we match the relevant personal information item to a place based on the shared related subjects.

OpenStreetMap Places. In the case of a place with a large geographic area such as a park, the matching can fail, as the point representing the place may not fall within the detected user's stay point. To this end, we leverage OpenStreetMap (OSM), an open-source database, to enhance the place matching process. When collecting the Foursquare places, we match them place with the corresponding place in OSM. In particular, we extract the polygon geometry information associated to the OSM place that corresponds to a given Foursquare place. For a given Foursquare place, we find a matching OSM place based on the location and name information. More precisely, we get all the OSM places with a polygon information that intersect with the geo-coordinates of a given Foursquare place within a 50-meter radius. We then perform a fuzzy match of the name of the Foursquare

place with the name of the OSM places (using the addrname and name fields). Using the places of Great Britain extracted from Foursquare, we managed to match 4.9% of the OpenStreetMap polygons with the venues. With the polygon information associated to the Foursquare venues, we match a stay point buffered with a 50-meter radius (that approximates the GPS accuracy) to a Foursquare place if they intersect with a polygon extracted from the corresponding OSM place.

User Places. For each user, we record all the places that the users have visited and for which they have further confirmed their visits. This creates a record of places visited by each user. It has been shown that people have a set of place that they are likely to visit again either as part of their routine (*e.g.*, home or work), or as part of their interests (typically, the third places such as pubs, supermarkets) [3, 30]. As so, from this record of user places, we learn the users' routine and the places they are likely to visit again, which helps improve our place matching algorithm. More specifically, we use the location and time information of the previous visits a user has made at the places in the vicinity of a detected stay point. Given the start and end times of the detected stay point, we retrieve the places within a 50-meter radius of the stay point whose visits fall within similar time ranges, with more or less one hour visit shift. The place matched is the one that has the greatest visit frequency.

Table 4. Personal information categories extracted from Foursquare and DBPedia.

Personal information categories	Foursquare and DBPedia place categories
Activities	Place categories related to **Entertainment** (*e.g.*, Venue), **Tourism** (*e.g.*, Monument / Landmark), **Education** (*e.g.*, University), **Sports** (*e.g.*, Stadium), **Eating Out** (*e.g.*, Restaurant), **Having a Drink** (*e.g.*, Bar), **Self-care** (*e.g.*, Nail Salon), **Shopping**, and **Accommodation** (*e.g.*, Hotel).
Addictions	All place categories related to **Smoking** (*e.g.*, Smoke Shop), **Alcohol** (*e.g.*, Bar and Pub), **Drugs** (*e.g.*, Marijuana Dispensary), and **Gambling** (*e.g.*, Casino).
Ethnic affinities	Restaurants with world cuisine such as French Restaurant and Mexican Restaurant, EthnicGroup.
Family life	Place categories associated to **Children** (*e.g.*, Elementary School, Middle School, Maternity Clinic, Baby Store, Playground, Toy / Game Store, and Day Care), **Wedding** (*e.g.*, Wedding Hall and Bridal Shop), and **Seniors & Retirement** (*e.g.*, Assisted Living).
Health	Place categories related to **Vegetarian** (*e.g.*, Vegetarian / Vegan Restaurant and Halal Restaurant), **Eating Disorders** (*e.g.*, Gluten-free Restaurant), **Injuries** (*e.g.*, Medical center, Emergency Room, and Hospital), **Surgery** (*e.g.*, Hospital and Emergency Room), **Drugs & Medication** (*e.g.*, Drug Store and Pharmacy), **Vision Care** (*e.g.*, Optical Shop), and **Oral & Dental Care** (*e.g.*, Dentist's Office).
Ownership	All place categories related to owning **Pets** (*e.g.*, Veterinarian), **Private Vehicles** (*e.g.*, Automotive Shop, Drive-in Theater, and Bike Shop).
Political influences	Voting Booth, Non-Profit, Town Hall, Organisation, Election, Non-ProfitOrganisation, PoliticalParty, MemberOfParliament, GovernmentAgency, TradeUnion.
Religious Beliefs	Places of the categories ReligiousBuilding, Diocese, ChristianBishop,Spiritual Center (*e.g.*, Church, Mosque, Synagogue) and restaurants of categories Kosher Restaurant, Jewish Restaurant and Halal Restaurant.
Sexuality and gender	All place categories determined from our ground truth dataset (see text for details) as well as places of the category Gay Bar.
Skills	All place categories related to **Higher education training** (*e.g.*, University), **Driving** (*e.g.*, Gas Station and Automotive Shop), **Arts** (*e.g.*, Art Studio and Photography Studio), and **Sports** (*e.g.*, Soccer Field).
Social activities	All place categories where people usually have social activities with **Friends**, **Family**, and **Colleagues** such as Restaurant, Bar, and Office.
Socio-economic status	All place categories determined from our ground truth dataset (see text for details).
Topics of interest	All place categories determined from our ground truth dataset (see text for details).