User Modeling and User-Adapted Interaction (2022) 32:25–90
https://doi.org/10.1007/s11257-021-09307-6

Transferring recommendations through privacy user models across domains

Frederic Raber[1] · **Antonio Krüger**[1]

Received: 4 June 2020 / Accepted in revised form: 24 September 2021 / Published online: 8 November 2021

Abstract

Although privacy settings are important not only for data privacy, but also to prevent hacking attacks like social engineering that depend on leaked private data, most users do not care about them. Research has tried to help users in setting their privacy settings by using some settings that have already been adapted by the user or individual factors like personality to predict the remaining settings. But in some cases, neither is available. However, the user might have already done privacy settings in another domain, for example, she already adapted the privacy settings on the smartphone, but not on her social network account. In this article, we investigate with the example of four domains (social network posts, location sharing, smartphone app permission settings and data of an intelligent retail store), whether and how precise privacy settings of a domain can be predicted across domains. We performed an *exploratory study* to examine *which* privacy settings of the aforementioned domains could be useful, and validated our findings in a *validation study*. Our results indicate that such an approach works with a prediction precision about 15%–20% better than random and a prediction without input coefficients. We identified clusters of domains that allow model transfer between their members, and discuss which kind of privacy settings (general or context-based) leads to a better prediction accuracy. Based on the results, we would like to conduct user studies to find out whether the prediction precision is perceived by users as a significant improvement over a "one-size-fits-all" solution, where every user is given the same privacy settings.

Keywords Privacy settings · User modeling · Recommender systems · Data privacy · Cross-domain user modeling

✉ Frederic Raber
frederic.raber@dfki.de

Antonio Krüger
krueger@dfki.de

[1] DFKI Saarland Informatics Campus, Saarland, Germany

1 Introduction

In the past, privacy settings were often neglected by users, especially when they perceive the shared data as non-harmful, like their posts in social networks (Majeski et al. 2011). Although users have become increasingly engaged in privacy settings, there is still some space for improvement (Dey et al. 2012; Stutzman et al. 2013). Studies have shown that even data inside a social network can be used for various attacks like stalking, identity theft, social engineering attacks, or face re-identification (Gross and Acquisti 2005). Users prefer more restrictive privacy policies; however, the default settings in social network sites are very permissive, significantly more permissive than desired by users (Watson et al. 2015). Social network sites and location services tried to tackle the problem by offering automatically generated friend lists, which allow an easy audience selection for new content to be published. However, studies have shown that only 17% of all posts are published using at least one of these automatic friend lists (Mondal et al. 2014), either because the additional effort is perceived as too high, or because the friend lists are not perceived as suitable for selecting the post audience (Mondal et al. 2014). Manually created friend lists typically lead to better results; unfortunately, they need even more user interaction and are thus not often used either (Paul et al. 2011). As a consequence, users have to pay attention when, for example, a photograph of them is taken which could be uploaded in social media, and apply workarounds to prevent being photographed and published on a website without their consent (Rashidi et al. 2018).

Research has tried to support users by automating the generation of privacy settings using machine learning, for example, by using privacy settings from earlier posts as an input or by asking the user for feedback on privacy settings (Sinha et al. 2013; Fang and LeFevre 2010; Lugano and Saariluoma 2007). Also, context factors like the occasion when sharing a location, or the recipients of a post or location, have been found to have an influence on privacy settings (Benisch 2011; Consolvo et al. 2005; Patil et al. 2012) and are threrefore taken into account for the prediction. Other approaches take the personality of the user and her general privacy preferences according to the IUIPC privacy questionnaire into account for deducing the privacy settings (Raber et al. 2017; Raber and Krüger 2018).

However, all those approaches need either domain knowledge (for example, privacy settings from earlier posts) from the domain where the privacy settings have to be predicted, or personality and privacy profiles of the user. In some cases, especially in the critical moment when a user has just created a new account and is overwhelmed by the privacy settings that he should adjust, none of the data is available, also known as the *cold start problem* (Schein et al. 2002). However, most users have already used other systems where privacy settings had to be selected which can be suitable as an input for the prediction, also known as *cross-domain user modeling*. Other models use ontologies to describe a user model, allowing to recommend the degree of information shown to the user based on the current user state and a reasoner (Heckmann 2006). Cross-domain recommender systems or ontology-based systems that are tailored especially towards privacy

settings have so far, to the best of our knowledge, not been part of research. In this article, we took four domains as an example, to find out whether and how well the privacy settings of a domain can be predicted using the privacy settings from one or several other domains.

As we have shown in earlier work, privacy decisions are not ultimately binary (Raber et al. 2017; Raber and Krüger 2018). In social network posts, a user might not just hide or show the complete post, but might want to take a middle road and hide only the post image or comments, while still sharing the post text with a certain friend group. Furthermore, different domains have different privacy options to set, which makes it hard to directly compare the privacy settings per se. We therefore use *privacy levels* in our prediction; these can be resolved to concrete privacy settings after a prediction, depending on the domain (see Sect. 3). In our work, we discuss two different granularities of privacy levels: first, *mean domain privacy levels* describe an average privacy level over all privacy levels for a user in a specific domain, independent of context factors. There is *exactly one mean domain privacy level* per domain. In contrast to that, there are multiple *context-based privacy levels* for a domain, one for each combination of each context factor instance (for example, there is one *context-based privacy level* for the post topic "family affairs" (context factor 1) in combination with the recipient group "school friends" (context factor 2). Currently, social media or location sharing services offer users a "one size fits all" solution for their users, where everything is set to a specific default value at the beginning. Using the user's *mean domain privacy level*, the service could already tailor all privacy settings to be more restrictive if the user has a high *mean domain privacy level* for that domain, or use a looser set of default privacy settings if the *mean domain privacy level* is low. Using the *context-based privacy levels*, one could tailor the privacy settings even better to the user, by providing different privacy settings for different contexts, for example, when a new post about "family affairs" has to be shared with "school friends". Privacy recommenders are often based on the same general privacy attitudes, for example, the IUIPC questionnaire (Malhotra et al. 2004). We therefore speculate that mean-based privacy levels and context-based privacy levels, although being specific to the domain, all contain a specific privacy attitude which is unique to the user and can therefore be used to transfer desired privacy settings to other domains.

In this article, we will build up on our previous work on deriving privacy settings using context factors and individual factors (such as user personality or privacy attitude). Based on the user studies presented in this article, we want to compare the four different domains of privacy levels for *social networks*, *location sharing, intelligent retail data*, and *mobile apps permissions* regarding how privacy levels for each of those domains can be predicted using the privacy levels from the other domains; and to what extent the usage of context-based privacy levels, using different values for each context factor, plays a role in this context. In contrast to previous work, we will not use solely context or individual factors as a source for the prediction. Instead, we will predict the privacy settings of a domain using the privacy settings from another of the aforementioned domains. In our work, we will investigate which of the two mentioned granularities, *mean domain privacy levels* or context-based privacy levels, work best for predicting the *mean domain privacy levels* and

the context-based privacy levels, which domains and, inside a domain, which of the privacy levels should be used for the prediction and how precise a prediction can be. For this purpose, we trained a cross-domain recommender system with a small training set, which can be seen as a lower bound for the precision that could be achieved using a large data set. According to our results, it is possible to use privacy levels from other domains for a prediction, but interestingly, the results also show that it is not always the case that using more fine-grained privacy levels, e.g., more data, leads to a better prediction. The approaches allow a prediction about 15–20% better than random, which looks small at a first glance. However, a traditional within-domain prediction, which forms an upper bound for the precision, can only lead to a precision of 20–25% better than random.

To conclude, the article has *two* main contributions: To the best of our knowledge, we discuss the first approach on *cross-domain privacy recommendations* on *four* exemplary domains. Furthermore, we compare *four* different techniques for the recommendation, involving either an average privacy level for a domain, or a multivariate set of context-based privacy levels for a domain.

In the next sections, we will first discuss related work, as well as our own previous work on predicting privacy levels using a user's personality and privacy attitude, before we present the user studies and results that we conducted in order to answer the aforementioned questions.

2 Related work

First approaches supporting users in choosing their privacy settings used existing privacy settings, for example, from social networks, to predict the remaining privacy settings on Facebook using machine learning. Sinha et al. used a Latent Dirichlet Allocation (LDA) and Maximum Entropy to propose privacy settings for a user's friend groups on Facebook (Sinha et al. 2013). Other approaches rely on user-generated input for the prediction of privacy settings, for which the user has to label some social network friends with privacy privileges. These partially complete settings are then used as an input to generate the privacy privileges for the remaining users (Fang and LeFevre 2010). Although such a supervised method in general produces more accurate results (Barnes 2006), it comes with an increased user burden. This is especially crucial, as most users need a trigger rather than an additional user burden, like social triggers where they interact with or observe other users, in order to become active (Das et al. 2019). Studies have shown that fully automated privacy recommenders are preferred to those that need user interaction (Zou et al. 2020) and that users also tend to abandom privacy recommendations if they are perceived as low-value or inconvenient (Zou et al. 2020). However, an automated prediction of privacy settings has always to be accompanied by salient privacy notices, especially about risky privacy practices (Ebert et al. 2021) or a privacy user interface that gives the user a quick overview on the current privacy state (Christin et al. 2013) and that allows to review and adapt the privacy or permission settings (Tsai et al. 2017). In the mobile phone domain, also shoulder surfing is a privacy issue which can be moderated by informing the user about such attacks (Zhou et al. 2016).

Some approaches even rely on a questionnaire that has to be filled out before the prediction can be used (Lugano and Saariluoma 2007). As a combination of both, there also exist semi-supervised methods that use existing social network profile information of the user and graph properties together with active learning methods to reduce the user burden (Shehab and Touati 2012). Using crowdsourcing for gathering training data for the prediction has also been discussed to find an effective tradeoff between usability of and data privacy (Ismail et al. 2015). Privacy recommenders using the permission type, app name and the current state of a smartphone app (i.e., which app is currently in the foreground and how visible the app sending the permission request is) can reduce the amount of unwanted disclosures by about 75% compared to the standard permission dialogue on Android (Wijesekera et al. 2018). Also data retention, i.e., obfuscating or deleting data after a certain time span defined by the user, is a privacy technique which is appreciated by users (Ebada Mohamed and Chiasson 2018), especially if they have the possibility to actively send data into retirement (Murillo et al. 2018).

Research also proposes the use of privacy stereotypes, which should simplify the process of choosing privacy settings: In the location sharing domain, most of the time, locations can be shared with only a few different sets of location sharing settings (Ravichandran et al. 2009). Only three privacy stereotypes match the user's privacy settings with an accuracy of 90% at any given time (Ravichandran et al. 2009). Another recent publication showed that it is possible to cluster users into five groups of users using a questionnaire, allowing to assign each of them a privacy policy which is tailored for their respective privacy needs (Lynn Dupree et al. 2016). This can also be done for the fitness domain using recommender systems and machine learning (Ref Sanchez et al. 2020). Privacy decisions can be predicted by what the authors call *cognitive heuristics* (Shyam Sundar et al. 2020), which are shortcuts that allow a fast decision-making for the user. If, for example, the website provider is a popular name, brand or organization, users often imply that the website provider can guarantee the security of the website and are therefore more willing to disclose private data. Lying about private questions online is also a widespread privacy protection behavior which can be predicted using the results from a privacy quesionnaire (Sannon et al. 2018).

Research on privacy stereotypes has shown that users often follow a multi-dimensional approach (Knijnenburg et al. 2013; Wisniewski et al. 2017) when deciding which data to share. This allows to distinguish users into user stereotypes based on their sharing behavior (Knijnenburg et al. 2013) or the used privacy strategy (Wisniewski et al. 2017). For example, one stereotype may be fine sharing location-related but not interest-related items, whereas another group may behave exactly the other way around. To which of the stereotypes a user belongs to, can be decided based, for example, on personality, or demographic factors (Knijnenburg et al. 2013). In our opinion, such individual and context factors like the data type or the personality of the user are one of many other (context) factors that play a role in the user's decision. In Sect. 3, we will point out some *exemplary* context factors which are present in different domains according to recent research, and which we will investigate in our work. However, which context factors influence the users' privacy decisions is an active research field, we therefore cannot use an exhaustive

list of context factors for our study. In our work, the context-based privacy levels form a multidimensional table of privacy decisions, accounting for the multidimensionality of privacy decisions in the aforementioned publications. We discuss both the recommendation of such multidimensional privacy levels (mean-based context-aware regression analysis (MCR) and context-factor-based context-aware regression analysis (CCR) method) as well as the usage of multidimensional privacy levels as a source for the prediction (context-factor-based regression analysis (CGR) and context-factor-based context-aware regression analysis (CCR) method).

The aforementioned approaches all predict privacy settings for a single domain, and are therefore also called *single-domain recommender systems*. If no or only sparse information is given about the user in the domain for which the recommendation has to be performed, those approaches fail. Cross-domain recommender systems use user models from several domains in order to derive recommendations even if the data about the user is unsufficent in the recommendation domain. Although single-domain recommender systems should be preferred due to their higher prediction precision, if available (Sahebi and Brusilovsky 2013), cross-domain recommender systems have the advantage that they are able to predict settings for more than one domain, leading to an increased user engagement and satisfaction (Adomavicius and Tuzhilin 2005).

Other approaches do *not* try to transfer the user model from one domain to another, but rather collect the user models from several domains in a common format like an ontology, so that an ontology reasoner can be used to infer a user model based on the data (Heckmann 2006). An example of such an approach is the Ubiquitous User Model by Heckmann (2006), consisting out of two parts: First a general user model (GUMO) containing general information about the user like demographic data, personality and characteristics, emotions, etc., together with domain-specific interests (like favorite movies or books), and second the Situation-Reports, which describe the current situation the user is in, according to sensor data from the Ubiquitous environment; for example, whether she is stressed according to the heart rate monitor, or whether she is in a hurry according to video cameras detecting the walking speed of the person. Based on these two concepts, the Ubiquitous User Model can give recommendations for a user interface, for example, that the navigation at the airport should be simplified if the user is in a hurry (Heckmann 2006).

Apart from approaches to predict privacy settings, researchers also found that context factors play a significant role in user's privacy decisions (Ebert et al. 2020), and identified different context factors that are important for the choice of privacy settings in different domains. In location sharing, several studies found that the person requesting the location is one of the main context factors (Benisch 2011; Consolvo et al. 2005). Some also state that the time and day of the week as well as the location plays a role (Benisch 2011). Also whether the person is in a relationship and in which stage of the relationship, plays a significant role on the sharing behavior with the user's partner (Young Park et al. 2018). Later studies reviewed these results and found that it is not the time and day that is the appropriate context factor (Patil et al. 2012), but besides the requestor, the user's occasion or activity is the second main context factor that is important for the decision whether to share

the location or not (Consolvo et al. 2005). The granularity of the shared location also plays an important role (Patil et al. 2012), meaning that the option to share a coarse-grained location like "only the city name" or "only the country" should also be available. Similar context factors could also be found in the social media domain, where the topic of the post, as well as the receiving friend of the post, are important (Raber et al. 2017). However, studies also found indications that the topic of the post plays only a minor role. Also the age of the user plays a role: Whereas younger users tend to decide based on trust, older users decide based on the perceived benefits for disclosing their data (Ghaiumy Anaraky et al. 2021). Another context factor found to be highly significant is whether users are paid for disclosing their data. Even if the negative consequences are clarified, people are likely to share their data when they are paid (Hutton et al. 2014). However, we do not want to support users being paid for disclosing more data than they desire, which led us to the decision to exclude this context factor for our research. In the mobile app domain, researchers achieved the best results when using the permission type and app id or category for the prediction (Liu et al. 2014, 2016), or a combination of app name, permission type, the foreground app, and the visibility of the app making the permission request (Wijesekera et al. 2018). Using a large database of about 4.8 million users, and those two context factors, a prediction accuracy of 64.28% to 87.8% is possible using machine learning (Liu et al. 2014). If user feedback is integrated, a similar semi-supervised approach was able to achieve an acceptance rate of 78.7% of the proposed settings. The domain of intelligent shopping data has been investigated by our previous research (Raber et al. 2018). In a study, we found the data type in question, as well as the requesting stakeholder (e.g., the retailer, third parties like marketing agencies, etc.) to have a significant influence on the privacy decisions.

All work presented here used different approaches to predict the privacy settings in a binary deny-or-allow fashion, using either existing privacy settings from the same domain, context factors that influence the privacy decision, or additional user input for their prediction. Other recommender systems instead use user behavioral data, for example, for developing user models for adaptive cybersecurity (Addae et al. 2019) or to infer a degree of diversity for recommender systems suitable to the user's personality (Wu et al. 2018). In the past, we already did some research on how the personality and privacy desires (according to the IUIPC[1] scale) can be used together with context factors as an input to predict privacy levels for specifying the correct audience for a social network post (Raber et al. 2017), the detail level of a shared location (Raber and Krüger 2018), which data out of an intelligent retail store should be shared with whom (Raber et al. 2018) and to assist the user in adapting the permission settings for her smartphone apps (Raber and Krüger 2017). Another approach of inferring privacy settings might be done using self-reported privacy measures as a basis for the prediction, allowing to better infer the user's actual privacy behavior rather than their reported privacy desires (Faklaris et al. 2019). Other researchers already explored whether form-based profiles (e.g., personal profile data entered into forms) and tag-based profiles (e.g., tagged photographs) can be inferred

[1] Internet Users' Information Privacy Concerns.

between different social web providers like Twitter, Facebook an Tumblr (Fabian et al. 2013). So far, to the best of our knowledge, it is unknown whether the privacy settings of a domain can be predicted using the privacy settings from multiple other domains, especially when considering privacy levels instead of binary privacy decisions.

3 Background

In earlier publications, we concentrated on deriving the privacy levels using what we call *context factors* (like the topic of a post or the occasion when a location is shared) together with *individual factors* (like the personality or privacy attitude of the user) to infer privacy levels for four different domains, namely for posts in a social network, location sharing services, the data recorded inside an intelligent retail store like Amazon Go[2], and the permissions of smartphone apps. Although individual factors are usually not available, both the big five personality traits, as well as the IUIPC privacy measures can be derived using data from the social web (Farnadi et al. 2016; Raber and Krüger 2018) without adding any user burden. As stated in the introduction, the disclosure decision is not ultimatively binary. Research has shown that when users express their privacy policy in a free-text form, they tend to have fine-grained privacy policies that also allow them to share an obfuscated position like the street or city center instead of the exact location (Patil et al. 2012). Recently, social network providers like Facebook also offer sharing an obfuscated location, for example, only the city of the current location, allowing the users to share their position without disclosing too much information. Inspired by the aforementioned work, our earlier studies also offered multiple privacy levels where it is technically possible.

Within all our studies, people actually used the offered fine-grained privacy levels. In the location sharing domain, for example, the intermediate privacy level "city only" was used most frequently, even before the two binary options "exact location" and "no location", indicating the user acceptance of this approach. Details on the actual frequency of use can be found in Raber et al. (2017), Raber and Krüger (2018).

For each of the four domains, we used a set of context factors, as well as the individual factors as an input for the prediction of the privacy levels. Table 1 summarizes the investigated domains, the used context factors and the privacy levels. Note that the input for each recommendation always consists of the listed context factors **and** the individual measures (IUIPC questionnaire and big five personality inventory). For a detailed description of the study and its outcome, please refer the following subsections.

Due to the privacy paradox (Barnes 2006), the user's online behavior when setting privacy settings significantly differs from their actual privacy desire. Therefore,

[2] https://www.amazon.com/b?node=16008589011.

Table 1 Overview of the earlier work on recommending privacy levels using individual measures and context factors

domain	Context factors	Privacy levels
Social media (Raber et al. 2017)	Post topic, friend group	Show on timeline, show only on page, hide images/comments, hide post, hide post and reshared post
Location sharing (Raber and Krüger 2018)	Occasion when shared, recipient	Exact location, street (and city) city only province only continent only no location
Mobile phone (Raber and Krüger 2017)	App category, permission type	Allow / deny
Intelligent shopping (Raber et al. 2018)	Data type, stakeholder	Depending on data type (see Sect. 3.4)

for all our studies, we investigate the user's privacy desire, i.e., the desired privacy settings using questionnaires rather than investigating their actual privacy behavior.

In this section, we will discuss our published research that is of importance for this article. Research described in later sections is new and unpublished so far.

3.1 Social network privacy levels

Social networks like Facebook currently only allow users to show or to hide a post based on the friend or friend group that is accessing the user's personal page. However, we found that besides the recipient group, the topic of the post is another context factor that influences the decision of the user (Raber et al. 2017). Therefore, we proposed five different privacy levels instead of a binary choice; this allows users to specify their privacy desires in a more fine-grained way. The privacy levels are shown in Table 1 (with ascending strictness from level 1 to 5).

To capture the individual factors, we used a custom questionnaire inspired by several privacy and personality questionnaires (Malhotra et al. 2004; Wisniewski et al. 2015) as well as some custom questions (Raber et al. 2017). We followed a two-step approach, using a "main study", for finding the significant items in our privacy questionnaire that allow the prediction of the privacy levels. For this purpose, we used a modified version of a wrapper-subset-selection (WSS) algorithm, and a ridge regression for the choice of questionnaire items. In the *validation study*, we tested the prediction in a realistic scenario where we asked participants to copy and paste actual Facebook posts out of their profile, along with the topic they would assign them to. Our ridge regression algorithm then proposed a privacy level for the different friend groups and asked the participants to correct any wrong predictions. On average, we achieved a root mean squared error (RMSE) of 0.74 for the prediction on our five-point scale, which is significantly better than a random prediction with an RMSE of

about 2.1. Moreover, for 144 out of 230 posts that were entered in the study, the participants preferred our predicted setting (without changes) to their own setting that they used when they published the post on Facebook. Both context factors, as well as the individual factors, played a significant role for the prediction precision.

3.2 Location-sharing privacy levels

The same problem also holds for a large portion of the location-sharing services: Mostly, recommenders for location-sharing permissions only offer a binary decision (hide/share). However, sometimes users want to give a rough indication of their current location without disclosing too much information. We therefore allowed users to select more fine-grained options by using the location abstraction levels from the Google Maps API (Raber and Krüger 2018). The detailed privacy levels are shown in Table 1 (with ascending strictness from privacy level 1 to 7).

As related work has shown, the most important context factors are the occasion when the location is shared, as well as the requesting person (Connelly et al. 2007), therefore we integrated those two factors into our study. As individual factors, we used the IUIPC (Malhotra et al. 2004) privacy questionnaire and the big five personality inventory (Costa and McCrae 1992) for capturing the privacy desires and personality of the participants. Using an online questionnaire, we captured the aforementioned individual factors as well as location-sharing settings for all combinations of context factors. According to an ANOVA analysis of our results, both the occasion as well as the requestor have a significant impact on the choice of the privacy level, whereby the occasion has the major influence ($F_{10,890} = 66.855$) compared to the requestor ($F_{8,1092} = 3.329$). These results go hand in hand with related work that also revealed the recipients and the occasion of the location sharing as important context factors for the sharing decision (Patil et al. 2012). Correlation analyses have shown that all measures of the IUIPC as well as the big five personality inventory have a high correlation with the chosen privacy level. Finally, using a categorical regression (CATREG) algorithm, we were able to predict the privacy levels with an apparent prediction error (APE) down to 0.80 (on a scale from 0 (=no error) to 1 (=same error as a random prediction) using all individual factors, showing that tailoring the privacy settings to the user's personality and privacy attitudes can again significantly increase the prediction precision. Integrating our approach based on individual features into a context-based approach from related work can therefore again increase the prediction precision.

3.3 Smartphone app permission settings

Unlike the previous domains, app permission settings (like the permission to access the GPS location, or access to contacts) have only two possible states - allow or deny - reducing the scope of possible privacy settings to only two, instead of five or seven different privacy levels. As context factors, we used the category of the app whose permissions have to be predicted (like "messenger" or "navigation app") as well as the permission type (for example, "use microphone" or "access GPS location").

Individual factors have been captured by the big five personality inventory and the IUIPC privacy questionnaire.

In our previous work (Raber and Krüger 2017), we concentrated on two different use-cases to assist the user in choosing his app permission settings. The first is a *permission wizard* that supports the scenario when a user has just bought a smartphone, and wants to set the permission settings for all apps in one go, based on his personality and privacy preferences. Second, we also want to support users that already have a smartphone with all permissions set, by supporting them in choosing their permission settings on a newly installed app in an interactive way. To be more precise, the approach observes the user traversing the list of app permissions, and takes action when the first permission setting is changed. The interactive approach then takes into account the permission settings up to this point, together with the context and individual factors, to predict the remaining permission settings of the app.

The results of our study showed that we were able to reduce the number of "clicks", i.e., the number of permission settings that the user had to adapt manually could be reduced by our interactive approach for about 25% of all cases. For 8% of the cases, the approach led to an increased number of clicks; for the remaining 67% the number of clicks remained the same. The permission wizard achieved up to 70% correct permission predictions using the IUIPC questionnaire together with the context factors as input.

3.4 Intelligent shopping data

In contrast to app settings, the data recorded inside an intelligent retail store does, for some of the data items, allow a more fine-grained setting than just to allow or deny the access (Raber et al. 2018). For example, the in-store movement trajectory can be blurred, or the access to billing data can be fully granted, or restricted to the total sum, the number of items bought, or the category of the items instead of the exact product ID. However, the possible options for the privacy levels are very different, as in the aforementioned examples of a movement pattern and billing data. Some of the data items, like personal data, allergies and nutrition preferences, can only be shared or unshared. In order to bring this set of highly diverse data into a common form, we used the different privacy options for each data item as a context factor (e.g., five different binary data items like "number of products bought", "total sales amount", "detailed receipt", etc., instead of one data item "billing information" with different granularity options), later called *binarized granularity*. As another context factor, we found the stakeholder that is interested in the data (e.g., the retailer, third parties, or family and friends) to be important (Raber et al. 2018).

As individual factors, we tested the big five personality (Costa and McCrae 1992) inventory and the IUIPC (Malhotra et al. 2004), as well as an additional custom questionnaire, for their suitability for predicting the privacy levels using a correlation analysis. However, the results indicate the personal inventory to be unsuitable for this kind of task. Using the IUIPC as well as the additional questionnaire in a follow-up validation study, we achieved a precision of up to 69% correct predictions with a ridge regression algorithm.

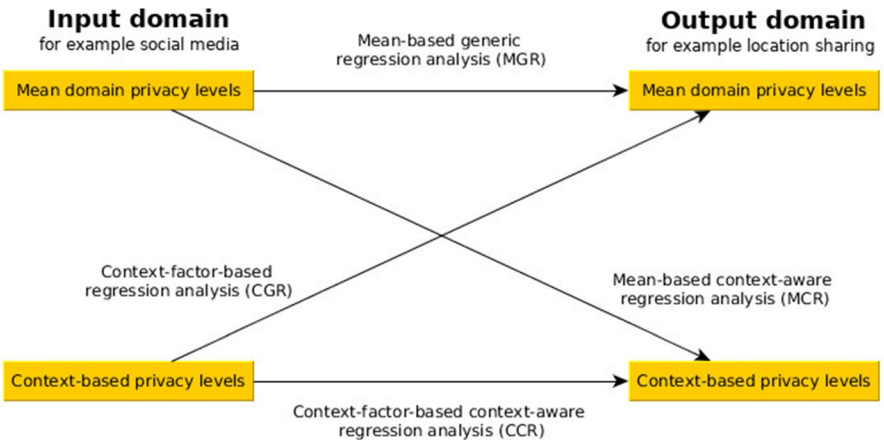

Fig. 1 Planned regression analyses

To conclude, we have shown that individual factors such as personality and privacy attitude play an important role in privacy decisions. But individual factors are often not available for a user, making it impossible to apply the aforementioned approaches. However, users have often already chosen privacy settings in other domains. In this article, we therefore investigate whether cross-domain prediction of privacy settings is possible.

4 Cross-domain user modeling for privacy settings

We performed two iterative studies using a bottom-up approach: In the first study ("exploratory study"), we performed regression analyses on the generic privacy levels for each domain and analyzed the regression coefficients,

1. To find out the domains containing privacy settings that are suitable regression coefficients ("input variables") for the regression of each of the four target domains
2. To identify which context-based privacy levels are potential candidates for improving the precision of the prediciton of

    (a) The generic privacy level of the target domain
    (b) The context-based privacy levels of the target domain

In the second study ("validation study"), we validate the choice of the context-based privacy levels using a fresh data set and compare the regression precision using either

(A) The generic privacy levels of the suitable domains as identified in 1 or
(B) The context-based privacy levels that have been identified as suitable in 2,

for predicting the generic privacy levels for each domain, as well as the context-based privacy levels.

In more detail, we pursue four different kinds of predictions, as depicted in Fig. 1. For the *mean-based regression analysis (MGR)*, we are working only on the top level, including the *mean domain privacy levels*, neglecting any context information, e.g., we are trying to predict *mean domain privacy levels* using the *mean domain privacy levels* of the other domains. For the context-factor based regression analysis (CGR), we are using context-based privacy levels on the input side, to predict the *mean domain privacy levels* on the output side. Conversely, the *mean domain privacy levels* could also be used as coefficients for a regression of the context-factor-based privacy levels (*Mean-based context-aware regression analysis (MCR)*). Finally, the context-factor-based context-aware regression analysis (CCR) uses context-factor-based input to predict context-factor-based privacy levels. Our ultimate goal is to find out which of the approaches works best for predicting the *mean domain privacy* and the *context-factor-based privacy levels*, which domains and, inside the domains, which context-factor-based privacy levels should be used, and what standard error can be achieved.

5 Exploratory study

The exploratory study and the validation study were conducted as an online study using a local LimeSurvey[3] installation at our institution. The participants were recruited using an online recruiting platform called Prolific Academic[4]. According to study results, the audience recruited through such a platform can be compared to an audience recruited through conventional methods like notice boards at a university or social network posts (Buhrmester et al. 2011). The participants were paid £2.10 for succesful participation, which needed 20 to 25 minutes, so that the minimum wage of £5 per hour for studies on prolific academic is guaranteed. At the beginning of the questionnaire, the participants had to confirm that they would carefully read and answer all questions, as not following the instructions or giving contradictory answers may lead to their participation being rejected. The actual questionnaire then asked demographic questions and whether the subject actively uses social networks and smartphones, and whether they know of intelligent retail stores like Amazon Go, offering a link to the Amazon Go teaser video on YouTube[5]. To ensure domain knowledge, we required them to be active users of at least one social network and to own and use a smartphone. When the participant entered responses not fulfilling these criteria, we ended the survey at this point. All other participants were then asked about their privacy preferences in the four aforementioned domains. The order of the domains was shuffled for each participant to avoid bias in the results.

[3] https://www.limesurvey.org.
[4] https://prolific.ac/.
[5] https://www.amazon.com/b?node=16008589011.

The exploratory and validation study had two different goals: Whereas the exploratory study aimed at finding and selecting suitable predictors for the recommender system, the validation study investigates the actual precision that can be achieved with a prediction, and how good the result is compared to a unparameterized method (i.e., a prediction without any inout coefficients) as a lower bound and a within-domain prediction as an upper bound.

At the beginning of each block of questions, we gave the participants an introduction to the situation that we are targeting, for example, "imagine you are creating a new social network post about your hobbies" for the social network domain. We then asked, for each group of recipients (like "friends", "family", etc.), which privacy option they would choose, according to the privacy options as presented in Section 3.1. This procedure was conducted for all combinations of

- Post topic and group of recipients (social network domain)
- Occasion and group of recipients (location-sharing domain)
- Stakeholder and data type (intelligent shopping domain)
- Application category and permission type (mobile phone domain)

To maintain comparability, we did not ask, for example, to rate whether the participant would allow or deny a permission of a specific app installed on her smartphone. Instead, we asked the participant, whether she would allow the permission for an app out of that app category in general. Therefore, no specific apps, user posts or shared locations were involved in the studies. As privacy options, we gave the users the same options that we gave the participants in earlier studies in the respective domains (see Sect. 3.1 for details), namely:

- Social network domain: five different privacy levels (show on timeline, show on page, hide images and comments, hide post, hide even if reshared)
- Location-sharing domain: the seven different location abstraction levels offered by the Google Maps API (exact location, street only, city only, province only, country only, continent only, no location)
- Intelligent shopping domain: allow/deny for each *binarized granularity* (see Sect. 3.1) option for all combinations of data type and stakeholder
- Mobile phone domain: allow/deny for each combination of app category and permission type

To assure the quality of the answers, we added four control questions in the section "location sharing" that the users had to answer exactly as stated in the task description. At the end of the questionnaire, the participants were offered a text box to enter any comments or ideas for improvements about the questionnaire. This procedure was reviewed and approved by the ethical review board of our institution.

5.1 Results and discussion

In total, 109 participants completed the questionnaire; they needed on average about 23 minutes for the task. Eight result sets had to be discarded as a control question was answered incorrectly, leading to 101 viable results. The age ranged from 18 to 64 years (mean 32.39, stdev. 9.55). Fifty-eight participants were female, 43 male. A gender effect could not be found within our results. 26 had already heard of intelligent retail stores, and 75 had not.

After importing the data for the analysis, we first computed several average values that will be used in the analysis later:

- For each context factor (see Sect. 3.1), we computed the mean privacy level for each instance of the context factor. For example, to compute the mean privacy level for the "events" occasion in the location-sharing domain, we averaged the privacy levels for the "events" occasion for the different recipient groups (e.g., we calculated the average over one column or one row in the context-based privacy levels table described in the background section). The averaged values will later be called (mean) context factor privacy levels.
- For each domain, we computed the average over all privacy levels, regardless of the context factors. These values are denoted as mean domain (privacy) levels.

We used the following procedure for the analysis:

1. *Context factor difference analysis* We performed a variance analysis on the *mean context factor privacy levels* for each domain and context factor, to find out which context factors lead to a significant difference in privacy levels.
2. *Mean-based generic regression analysis* We performed a regression analysis on the *mean domain privacy levels* for each domain, using the *mean domain privacy levels* of the other domains as regression coefficients, to determine which other domains are of influence for a prediction of the privacy levels.
3. *Mean-based context-aware regression analysis* We performed the same kind of analysis on the *mean domain privacy levels* for each domain, using the context-based privacy levels of the other domains as regression coefficients, to determine how precise a prediction on the context-based privacy levels can be when using only *mean domain privacy levels*.
4. *Context-factor-based generic regression analysis* Conversely, we performed a regression analysis on the *mean domain privacy levels* using the *mean context factor privacy levels* as regression coefficients, to determine which context factor instances could be of influence for the prediction. At this point, we were only interested in filtering out potential candidates, and building up hypotheses on which context factor instances could be of influence. As we were reusing the same data for multiple analyses, the reported significance values cannot be used to determine which context factor instances are significant without applying alpha correction. We therefore validate the results of the exploratory study later in the validation study using a fresh data set.

Table 2 Average privacy levels of our 109 participants and tests for variance on the context factors using Cochran's Q or Friedman tests: All context factors apart from the permission type in the mobile phone domain have a significant influence on the privacy levels

Domain	Avg. privacy level	Context factor	Statistic	Asymp. sig.
Social	2.05	Topic	263	< 0.001
		Friend group	254	< 0.001
Location	3.91	Occasion	154	< 0.001
		Requestor	328	< 0.001
Mobile	0.57	Category	53	< 0.001
		Permission	9	0.08
Shopping	0.46	Data type	74	< 0.001
		Stakeholder	54	< 0.001

Privacy levels in the social domain range from one to five, in the location domain from one to seven, and in the mobile and shopping domain from zero to one; bigger values mean more restrictive privacy levels

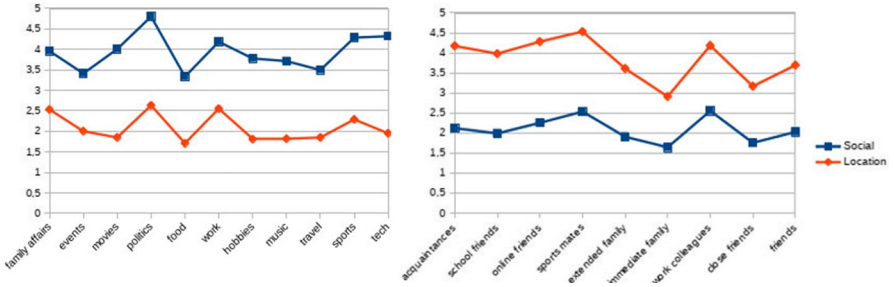

Fig. 2 Average privacy levels for the context factor instances of the topic/occasion (left) and friend group/recipient (right) context factors in the social media and location-sharing domain

5. *Context-based context-aware regression analysis* Finally, we performed a regression analysis on the *context-based privacy levels* for each domain, using the *context-based privacy levels* of the other domains as regression coefficients, to determine how precise a prediction on the context-based privacy levels can be when using *context-based privacy levels*.

The results of the above analyses will be described in the next subsections

5.1.1 Context factor difference analysis

Prior to the analysis, we tested each data set for normal distribution and sphericity, to decide on the correct statistical test for the variance analysis (e.g., ANOVA or its non-parametric equivalent, the Friedman test). For each domain and context factor, we had at least one context factor instance, for which the mean context factor privacy levels were *not* normally distributed. Therefore, we performed Cochrans's Q test for the binary scales (mobile phone and shopping domains) and Friedman tests

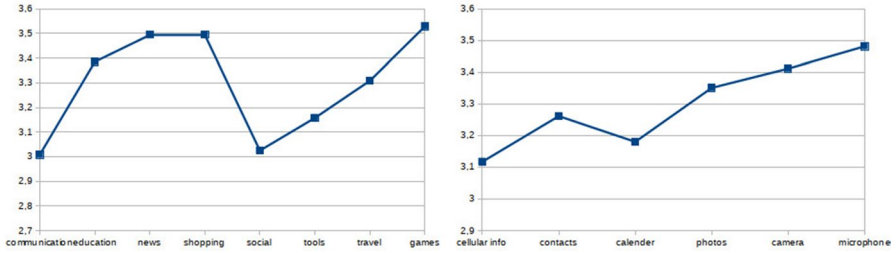

Fig. 3 Average privacy levels for the context factor instances of the app category (left) and permission type (right) context factors in the mobile phone domain

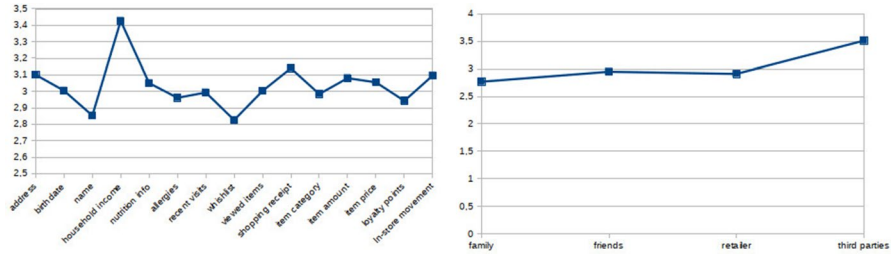

Fig. 4 Average privacy levels for the context factor instances of the data type (left) and stakeholder (right) context factors in the shopping domain

for the two non-binary domains (social media and location sharing). In addition to the variance analysis, we computed the mean privacy levels for each domain. As the scales have a different size, we added a normalized percentual average privacy level for each entry. The results for the analysis are shown in Table 2. The average values for each context factor instance are shown in Fig. 2 for the social and location domain, in Fig. 3 for the mobile and in Fig. 4 for the shopping domain.

For most of the context factors that we chose according to related literature, we confirmed their significant impact on the choice of privacy settings in the *context factor difference analysis*. Detailed results can be found in Table 2. Especially for the social network domain, both the recipient group (or friend group) as well as the topic of the post are very important context factors. For the location-sharing domain, the requestor seems to have a larger impact on the privacy setting than the actual occasion, which supports earlier work (Benisch 2011; Consolvo et al. 2005) which came to the same conclusion, that the requestor and occasion are the most important factors when the location is shared. Whether the requestor or occasion is more important, differs between earlier publications: Some see the requestor to be the most important factor (Benisch 2011; Consolvo et al. 2005) whereas others found the occasion to have a higher influence (Raber and Krüger 2018). Due to our results in the intelligent shopping domain, the stakeholder requesting the data is also a significant context factor, but is less important than the data type (for example, viewed products or the in-store movements) that is requested. We assume that the high diversity of data in the intelligent shopping domain might cause this effect, as the

need for privacy differs more for data types like household income or in-store movements, which might be considered more private than one's birthday or loyalty points earned throughout the shopping processes. However, as other studies on the importance of context factors also led to different results in other domains, the results have to be validated in further studies. In the location-sharing domain, the data type is always the same, and therefore yields a similar perceived criticality when shared unintentionally. This leads to the assumption that the importance of the "data type" as a context factor might rely on the diversity or number of data types (for example, whether data in the domain consists only out of gps locations, or whether there is demographic data, financial data and location data inside the domain), or both, which should be further investigated in future research. Another aspect that could have an impact on the perceived criticality could be the ability of the user to enter fake data. A user's birthdate, for example, can rarely be validated by the requestor of the information and is prone to be filled in with fake data, and could therefore be perceived as less critical. Interestingly, although the current privacy user interfaces in smartphone operating systems (Android or iOS) are tailored towards setting the permission individually for each permission type, our results indicate that the difference in privacy levels between permissions is not significant. In contrast to that, the category to which the app belongs has a strong influence on the privacy level. It seems that either the users trust apps from a certain category and grant the permissions, or they do *not* trust that *kind* of app and deny all of them. If the results can be supported by future studies, smartphone suppliers might want to redesign their permission UI, and include the app category as an option to let the user decide whether an app from that category should receive all permissions, or whether only some of them should be granted. The average privacy levels for location sharing and mobile apps are higher than for the other two domains, signaling that both location sharing and mobile app permission settings are perceived as more critical, or the recipient groups less trustworthy than for the two other domains. Apart from the permission type in the mobile app domain, all context factors have been proven to have a highly significant impact on the permission settings, supporting earlier work that relies on context factors for recommending privacy settings (Patil et al. 2012). The average privacy levels for location sharing and mobile apps are very similar when normalized to a percentual scale, i.e., an interval [0;1], resulting in 56% and 57%, respectively, and higher than the mean privacy levels for social media (M=2.05, normalized 41%) and shopping (M=0.46, normalized 46%), indicating that user tend to share less on social media and location-sharing services compared to mobile phone apps and intelligent retail shops.

5.1.2 Mean-based generic regression analysis (MGR) and Context-factor-based generic regression analysis (CGR)

Mean-based generic regression analysis (MGR)

To find out which domains are reasonable regression coefficients for a certain domain X, we first performed a separate regression analysis for each domain different from X, followed by a regression analysis including all domains that have a tendency to become significant coefficients (meaning $p < 0.10$). As an algorithm

Table 3 Regression analyses for the mean domain privacy levels

Target domain	Coefficients	R^2	$Adj.R^2$	Stderr	F	Sig.
Location	All	0.25	0.23	1.41	11.57	< **0.001**
	Social	0.21	0.20	1.43	28.65	< **0.001**
	mobile	0.03	0.02	1.59	3.02	**0.09**
	Shopping	0.10	0.09	1.53	11.83	< **0.001**
	Social+mobile	0.22	0.20	1.44	14.56	< **0.001**
	Social+shopping	0.25	0.23	1.41	17.40	< **0.001**
	Mobile+shopping	0.10	0.08	1.54	5.86	< **0.001**
Social	All	0.24	0.22	0.63	11.13	< **0.001**
	Location	0.21	0.20	0.63	28.65	< **0.001**
	Mobile	0.05	0.04	0.70	5.47	**0.02**
	Shopping	0.09	0.08	0.68	9.97	< **0.001**
	Location+mobile	0.23	0.22	0.63	16.07	< **0.001**
	Location+shopping	0.24	0.22	0.63	16.30	< **0.001**
	Mobile+shopping	0.09	0.07	0.68	5.30	**0.01**
Mobile	All	0.20	0.18	0.44	8.93	< **0.001**
	Location	0.01	0.01	0.49	1.58	0.21
	Social	0.04	0.03	0.48	4.91	**0.03**
	Shopping	0.19	0.19	0.44	25.54	< **0.001**
	Location+social	0.04	0.03	0.48	2.48	0.09
	Location+shopping	0.19	0.18	0.45	12.68	< **0.001**
	Social+shopping	0.20	0.18	0.44	13.25	< **0.001**
Shopping	All	0.32	0.30	0.42	16.49	< **0.001**
	Location	0.06	0.05	0.49	6.45	**0.01**
	Social	0.06	0.05	0.49	6.51	**0.01**
	Mobile	0.29	0.29	0.42	44.44	< **0.001**
	Location+social	0.08	0.06	0.49	4.50	**0.01**
	Location+mobile	0.32	0.30	0.42	24.51	< **0.001**
	Social+mobile	0.31	0.30	0.42	23.66	< **0.001**

All significant regessions are printed in bold face

for the regression analysis, we picked the ordinary least squares (OLS) algorithm as it has been shown to be robust also for data which is not normally distributed (Theil 1971). The OLS algorithm has been used for all statistical regression analyses within this article. For each regression, we report the goodness of fit (R^2) and adjusted goodness of fit ($adj.R^2$) describing how well the regression curve fits the data, as well as the results (F and significance) of the ANOVA analysis, describing whether a prediction using a regression produces viable results with the given coefficients. The results are shown in Table 3. Note that it is typical for a regression that measures like R^2 and results of the variance analysis are the same for a regression on X using Y as a coefficent, as they are for Y using X as a regression coefficient. However, to maintain readabilty and an easy comparison of the coefficients,

we included both combinations in the table; therefore some entries may seem to be duplicates. All significant coefficients according to Fisher (Fisher 1971) are printed in **bold** face.

The best regression coefficients for the location domain are the *mean domain privacy levels* from the social and shopping domains, which result in highly significant predictions. However, the mean privacy level of the mobile domain is very low (R^2); the ANOVA further implies that the prediction does not generate viable results, leading to the assumption that this domain is not of use for the prediction of a mean location-sharing privacy level. Therefore, using the coefficients from social and shopping allows a prediction as good as one using the coefficients from all domains. As a comparison, a prediction of the privacy level without using input variable (i.e., the standard error of the means of the location privacy levels) would lead to a standard error of 1.61. A similar picture can be seen for the social network domain, where the best coefficient is the *mean domain privacy level* of the location domain. However, the shopping domain as a coefficient still produces viable results, as well as a regression using the mobile app category privacy level. Also here, using location and shopping leads to a goodness of fit and standard error equal to using all domains as an input. A prediction without any input would result in a standard error of 0.71.

Having a look at the nested analyses (i.e.,) where exactly *two* other domains have been used for a prediction, we can see that in the social, mobile phone and shopping domain, adding a second domain to the domain with the lowest standard error does not significantly decrease the standard error. In the location-sharing domain, social + shopping is the combination of the two domains which have been found to be significant, forming the sources for the "all" predictor in that domain. Other combinations for that domain did again not lead to a significant improvement. To conclude, we can see that using nested inputs does not significantly increase the prediction over single-input predictions, or they are equivalent to the "all" predictor (in the location-sharing domain). We therefore did not further investigate the usage of nested predictions in the validation study.

The generic mobile app permission privacy level can be predicted best using the *mean domain privacy level* from the shopping domain. The *mean domain privacy level* from the social network domain is also a viable coefficient, whereas the *mean domain privacy level* from the location-sharing domain is of no use for this kind of prediction. Combining the input of the shopping domain with the coefficients of other domains could not lead to an improvement of the standard error. Predicting the mean privacy levels without any input would have a standard error of 1.01. Lastly, the *mean domain privacy level* from the shopping domain is predicted best using the mean privacy level of the mobile app domain, followed by the *mean domain privacy level* from the social and location domains which both provide a viable prediction. Also here, using only the coefficients from the mobile phone domain leads to the same standard error as combinations of the mobile phone domain and other domains. A prediction without any input would lead to a standard error of 0.92. The results will be discussed together with the results of the CGR method in the next paragraph.

Table 4 Tendentially significant regression coefficients (context factor instances) for the prediction of the domain privacy levels

Target domain	Coefficients	Instance	t	Sig.
Location	Social - topic	Events	2.49	0.015
		Movies	1.753	0.083
	Social - recipients	School friends	1.661	0.099
	Mobile - category	–	–	–
	Shopping - data type	Amount	1.672	0.098
	Shopping - stakeholder	–	–	–
Social	Mobile - category	Games	1.696	0.093
	Shopping - data type	–	–	–
	Shopping - stakeholder	–	–	–
	Location - occasion	Food	1.79	0.077
		Travel	2.071	0.041
		Tech events	2.203	0.03
	Location - requestor	Immediate family	1.804	0.075
		Extended family	1.803	0.075
Mobile	Social - topic	Sports	2.215	0.029
	Social - recipients	Close friends	2.008	0.048
	Location - occasion	–	–	-
	Location - requestor	Extended family	1.855	0.067
		Close friends	1.668	0.099
	Chopping - data type	Birthdate	1.78	0.077
		Income	1.88	0.064
	Shopping - stakeholder	Third parties	3.818	< 0.001
Shopping	Social - topic	–	–	–
	Social - recipients	Immediate family	1.959	0.053
	Location - occasion	–	–	–
	Location - requestor	Immediate family	2.067	0.042
	Mobile - category	Social media	1.964	0.052

The listed coefficients will later be used as an inputfor the CGR and CCR predictions.

Context-factor-based generic regression analysis (CGR)

For the context-factor-based generic regression analysis, we used the *mean context factor privacy levels* as regression coefficients to find out whether an increased detail level (e.g., one privacy level for each combination of context factors instead of one generic domain privacy level) can lead to an increased prediction precision in the regression. For this purpose, we first had to find out which instances of each context factor are suitable coefficients. However, as stated earlier, the significance values here cannot be seen as final (without using alpha correction or validation in a follow-up study), as the same data set is used multiple times. Later, we will validate the choice of context factor instances in the validation study using a fresh data

set. Similar to the analyses described above, we select all context factor instances that have the *tendency* to become significant ($p < 0.1$), so that we do not omit any instance that might be significant within another data set, while eliminating other instances that will most likely not become significant and that would disturb the regression algorithm. Note that we excluded the "permission type" context factor, as it was found to be insignificant in the context factor difference analysis. The results can be found in Table 4. A selection of scatter plots of the coefficients can be found in Fig. 5 in the Appendix. For a complete list of context factor instances and their average values, please refer to Figs. 2, 3 and 4.

Similar to the results of the generic regression analysis, the domains with the lowest standard error and the highest precision in the generic regression analysis have the highest number of significant context factor instances here. The location domain can be predicted best by the context factor instances from the social media domain, namely the privacy level for posts about events and movies as well as the privacy level for the friend group "school friends". However, the privacy level for the number of products bought in the shopping domain can also be used as a regression coefficient.

A similar picture can be seen for the location-sharing domain, which can be predicted best using the context factor privacy levels from the location-sharing domain, where the privacy levels of the occasions about having "food", "traveling" and "tech events" are found to be suitable, together with the privacy level of the requestor groups "immediate family" and "extended family". From the other coefficients, only the "games" category of the mobile app domain had a tendency to be a useful regression coefficient.

For the mobile app domain, we found coefficients from different domains to be useful. Most are found in the intelligent shopping domain, namely the privacy levels of the "income" of the customer and his "birthdate", as well as, with high significance, the stakeholder "third parties". Using the regression coefficients from the location-sharing domain, only the context factor privacy levels of two requestors, namely "close friends" and "extended family" seem to be suitable. From the social media domain, the "sports" topic and the recipient group "close friends" are both statistically significant regression coefficients.

Lastly, the shopping domain can only be predicted by few coefficients from different domains. From the social media domain, the recipient group "immediate family" is found to be suitable. The same coefficient "immediate family" is a significant coefficient from the context factor "requestor" from the location-sharing domain. From the mobile app coefficients, the privacy level for "social media" apps seems to be suitable .

To conclude the results of the MGR and MCR methods, we have two clusters of domains that can profit from each other for a prediction using a regression: The location-sharing domain and social media domain privacy levels seem to be good regression coefficients for each other and form the *"location-social cluster"* on the one hand, whereas mobile app settings and intelligent shopping privacy levels form another cluster, later called *"shopping-mobile cluster"*, that allow a good prediction of each other's privacy levels. For location-sharing and social media, if the data of the other domain is not available, the privacy levels from the shopping domain and

to some extent also from the mobile app domain (for social media) can be used. But if the data of the other domain inside the cluster is available, adding the coefficients from the other domains does not reduce the standard error for the location-social cluster. So if the *mean domain privacy level* of the other domain inside the cluster is available, the privacy levels from the other domains can be omitted according to our results.

For the shopping-mobile cluster, the situation is not that clear. Although using the *mean domain privacy level* of the cluster partner as a coefficient leads to the lowest standard error compared to the other domains, the social and location privacy levels are very good alternatives, especially for the intelligent shopping domain, where both coefficients are found to be significant. Therefore, combining all three domains leads to the lowest standard error for the intelligent shopping domain, although the two additional domains outside the cluster reduce the standard error only from 0.31 to 0.30. For the mobile app domain, adding coefficients other than the mean privacy level of the cluster partner cannot improve the prediction precision; therefore, if the privacy level from the intelligent shopping domain is available, all other data should be omitted for best results. If this is not the case, the privacy level from the social media domain also allows a prediction slightly better than random. On the other hand, data from the location-sharing domain is useless and should not be used for this domain.

Interestingly, the two clusters always contain the two domains that have a similar granularity, meaning they have a similar number of privacy levels (see Sect. 3.1). Whereas it seems clear that it is hard to use a binary scale from the shopping-mobile cluster to predict a more fine-grained scale from either the social media or the location-sharing domain, this should not be the case for the other way around. However, the context-based location and social privacy levels were also of no use to enhance the regression for the mobile app domain. Also for the intelligent shopping domain, the decrease in standard error is small. We therefore suppose that the existence of those two clusters is not a main product of the difference in their scales, but it is caused by some other factor, like the type of occasion when the decision is made (for example, whether it is made incidentally on the go for mobile apps or during shopping inside an intelligent retail store vs. as a main task during a leisure activity for the two others) or the type of privacy (privacy against companies like app manufactures or retailers on one hand, and friends or family members on the other). Which factors finally led to the clustering of domains, and which other clusters exist, should be further investigated in future work.

5.1.3 Mean-based context-aware regression analysis (MCR) and context-based context-aware regression analysis (CCR)

Both analyses try to predict the *fine-grained* context-based privacy levels. The mean-based context-aware regression analysis uses *coarse-grained* mean domain privacy levels as a source for the prediction, whereas the context-based context-aware regression analysis uses *fine-grained* context-based privacy levels as an input. The procedure used here was the same that we employed for the CGR method, e.g., we performed a regression analysis on all input variables. To exclude all input variables

where an impact on the outcome should not be expected according to Fisher (Fisher 1971), we selected those with a p-value < 0.1 for the validation study. Taking all four domains into account, we have a total of about 100 context-based privacy levels, with up to four viable coefficients each for the MCR, and again up to about 100 coefficients for the CCR. For each of those combinations, we would have to report t and significance values. For the sake of brevity, we will not report and discuss all viable coefficients here, but report the results of the final regression analysis in the validation study in Tables 9, 10, 11 and 12.

In general, we can see that, similar to the generic regression analysis, the domains inside the corresponding cluster in general produce the most, and most significant, context-factor-based coefficients (CFB-coefficients), supporting the correctness of the analysis. Having a look at the most significant CFB-coefficients (Table 4), we can see that the coefficient "third parties" in the "mobile apps" target domain has the highest significance of all of them. However, other stakeholders, like the retailer, do not even have the tendency to become significant ($p > 0.1$). This fact leads to the assumption that the trust that consumers put in app manufacturers is comparable to the trust they put in third parties like marketing companies, and not like the trust they put in a retailer. This is interesting, as app manufacturers and retailers are both the direct providers of the service the customer requests, unlike marketing companies, which usually do not offer a direct advantage for the customer. A possible explanation for this circumstance might be that the mere size, brand awareness or privacy image of a company is a key indicator for the trust in terms of privacy, rather than the service quality or the benefit from the service. This assumption is supported by other significant CFB-coefficients of the shopping domain, like one's birthdate and household income. Both data types are on average perceived as very sensitive by most customers, and therefore are shared rarely, indicating that the trust in mobile app developers and the will to offer them access to app permissions is relatively low. Interestingly, the other CFB-coefficients (outside the location-social cluster) that have been found useful for the mobile app domain indicate a less privacy-sensitive behavior. Posts about sports and for close friends or the extended family are usually not very restricted. However, having a look at the results of the generic regression analysis, including these other coefficients in the prediction actually reduces the precision and increases the standard error; we therefore assume that they are just statistical artifacts.

For predicting the social media privacy level, both the location-sharing settings from the immediate and extended family are found to be useful, meaning that the privacy levels used when sharing the location with members of the family are similar to those used in social networks. Having a look at the mean domain privacy levels, we can see that location-sharing privacy settings are typically stricter than settings for social media posts. Furthermore, users usually have relatively loose privacy settings for their family members (Raber et al. 2017). However, as the location-sharing domain is stricter in general, we assume that the loose settings in this stricter domain can be compared to an average privacy level in the social media domain, making it a good regression coefficient. Similar to this, more private occasions like traveling or preparing meals together ("food") have been found to be good CFB-coefficients.

For the location-sharing domain, most viable coefficients have their origin in the social media domain. The best coefficient is the privacy level of the posts with topic "events", most likely because events are the occasion where users typically share a location. Also "movies", e.g., social network posts about watching movies together or going to the cinema, are common occasions when a location is shared, making it the second most important CFB coefficient. We also found the recipient group (friend group) "school friends" to be viable, although the p-value is relatively high. In our opinion, locations are usually shared when a user does something interesting in her life, like attending events or having a meal at an expensive restaurant. The things in your life that you want to tell your school friends in order to improve your image are typically the same things. The same might hold for the amount of items that you bought at a shop. Therefore we think both privacy levels correlate to the location-sharing domain privacy level, and hence are good CFB-coefficients for the regression.

Lastly, in the shopping domain, we have one CFB-coefficient from each of the other domains. The results indicate that the recipient or friend group "immediate family" both from the social media and location-sharing domains, as well as permission settings for social media apps, are good coefficients. As stated before, most users use relatively loose privacy settings for their immediate family. Furthermore, social media apps require a lot of different permissions in order to be fully functional, like access to stored images or the location, which may lead users to grant them these permissions. So overall, the shopping domain seems to be a domain where customers feel confident when sharing their data, because they either do not see much harm in oversharing them, or because their trust in retailers is relatively high. The lower average domain privacy level supports this assumption. In this section, we investigated which input variables should be used for the different methods. In the next section, we will validate the choice of the aforementioned regression coefficients (Tables 4 and 3 and the candidates from the MCR and CCR method) and compare the prediction precision between the four methods in a validation study using a fresh data set.

6 Validation study

In the exploratory study, we had the goal to get a first impression of how accurate the prediction of the mean domain privacy levels can be when the mean privacy levels of the other domains are used, which domains are useful for a prediction, and which context factors (like the recipient of the data or the occasion, see Sect. 3.1) and mean domain privacy levels (e.g., the average over all privacy levels of a domain, see Sect. 4) are potential candidates for the prediction of the domain and context-based privacy levels. In the validation study, we will validate the results from the exploratory study, especially how well the regression performs with the selected coefficients, and which of the outlined approaches (MGR or CGR for predicting mean domain privacy levels, and MCR or CCR for the context-based privacy levels) performs best. For all four approaches, we will perform a statistical analysis to determine whether the regressor delivers

significant predictions, as well as a machine learning analysis in order to find out how precise the prediction would be if it was implemented into a recommender system, and what users could expect from the system. For a clearer understanding on the precision quality, we will report how good a prediction would be without using any input coefficients as a lower bound (typically by reporting the standard error of means, SEM, of the variable to be predicted), as well as a within-domain prediction for the CGR and CCR methods as an upper bound for the prediction precision.

As we found out in the context factor difference analysis, the permission type does not have a significant influence on the permission level. We therefore excluded this context factor in the validation study. Apart from this change, the procedure for collecting data was similar to that of the exploratory study.

We had 117 participants in the validation study, out of which 11 were discarded, as they answered a control question incorrectly, resulting in 106 valid records. The participants were aged between 18 and 71 years (average 32.67) and needed on average about 27 minutes for the task. 52 participants were female, 54 male. A gender effect could not be found within our results. All of them use a smartphone, 38 had already heard of intelligent retail stores; 68 had not.

6.1 Results

In the validation, we again performed the regression analyses from the validation study with the newly collected data set. Instead of using all coefficents for the regression, we used exactly the regression coefficients that were found viable in the exploratory study. To have an idea on how accurate a prediction can be using a machine learning algorithm, we also performed a machine learning analysis using a SVR (where multiple discrete privacy levels are predicted) and SVC (where a binary choice is predicted) algorithm. For the analyses involving a SVR, we report the mean squared error (MSE), whereas the area under curve (AUC) and F1-values are reported for the SVC analyses. To be more precise, we again performed:

- *Mean-based generic regression analysis (MGR)* trying to predict mean domain privacy levels from a domain using mean domain privacy levels from other domains
- *Context-factor-based generic regression analysis (CGR)* trying to predict mean domain privacy levels from a domain using context-based privacy levels from other domains
- *Mean-based context-aware regression analysis (MCR)* trying to predict context-based privacy levels from a domain using mean domain privacy levels from other domains
- *Context-factor-based context-aware regression analysis (CCR)* trying to predict context-based privacy levels from a domain using context-based privacy levels from other domains

Table 5 Validation study results for the MGR analysis

Target domain	Coefficients	MSE	R^2	$adj.R^2$	Stderr	F	Sig.
Location	All	**1.75**	0.32	0.30	1.21	14.43	< 0.001
	Social	**1.82**	0.23	0.22	1.27	38.01	< 0.001
	Mobile	1.85	0.14	0.14	1.34	21.37	< 0.001
	Shopping	1.94	0.15	0.14	1.34	21.89	< 0.001
Social	All	0.32	0.27	0.26	0.50	15.69	< 0.001
	Location	**0.28**	0.23	0.22	0.51	38.01	< 0.001
	Mobile	0.35	0.07	0.06	0.56	9.86	< 0.001
	Shopping	0.31	0.14	0.13	0.54	20.73	< 0.001

The best source domain for the prediction and the best precision score are printed in bold. Predictions are about 15%-20% better than random

Table 6 Validation study results for the MGR analysis continued

Target domain	Coefficients	AUC	F1	R^2	$adj.R^2$	Stderr	F	Sig.
Mobile	All	**0.70**	0.53	0.16	0.15	0.46	12.00	< 0.001
	Social	0.61	0.38	0.05	0.04	0.49	6.00	0.02
	Location	**0.69**	0.52	0.12	0.11	0.47	17.39	< 0.001
	Shopping	0.67	0.51	0.16	0.15	0.46	23.32	< 0.001
Shopping	All	**0.67**	0.46	0.15	0.13	0.46	8.68	< 0.001
	Social	0.57	0.45	0.09	0.08	0.47	12.24	< 0.001
	Location	0.66	0.46	0.06	0.05	0.48	8.04	< 0.001
	Mobile	**0.67**	0.55	0.10	0.10	0.46	14.71	< 0.001

The best source domain for the prediction and the best precision score are printed in bold

In the following section, we will first present the results from the statistical and machine learning analyses. The interpretation of the results follows in the discussion section later.

6.1.1 Mean-based generic regression analysis (MGR)

As stated before, we use the mean domain privacy levels from the exploratory study that have been found to be viable (e.g., $p < 0.1$). The results can be found in Tables 5 and 6. The machine learning results for the domains where a privacy level had to be predicted using a regression algorithm are shown together with the mean squared error (MSE) in Table 5, whereas the domains where a binary decision had to be predicted are shown in Table 6 together with the area under curve (AUC) and F1-value, which is typical for describing the prediction precision for these kinds of machine learning task.

Remember that the MGR is based solely on mean domain privacy levels, meaning it uses mean domain privacy levels from other domains as an input to

predict the *mean domain privacy level* of the target domain. Context factors or context-based privacy levels are not used.

For the location-sharing domain, the social media mean privacy level allows the most precise prediction (stderr=1.27). Both the shopping (stderr=1.34) and the mobile (stderr=1.34) mean privacy level result in a higher standard error in the analysis. A regression model without any input coefficients (e.g., the standard error of the means, SEM) would lead to a SEM of 1.44 within the validation study data, which is notably higher than the standard error of our regression model. Compared to the experimental analysis, the precision and R^2 values of the value pair "location sharing" – "mobile apps" have significantly improved. In the social media domain, including all other mean domain privacy levels led to a standard error of 0.50 on the seven-point privacy level scale for this domain. Using only single mean domain privacy levels as an input, the location domain performs best (stderr=0.51), followed by the shopping (stderr=0.54) and mobile (stderr=0.56) privacy levels. The SEM of the social media domain is 0.58 and therefore higher than any of the regression anlyses for this domain. The mean privacy level of the mobile app domain can be predicted best by using either the mean privacy level of the intelligent shopping (stderr=0.46) or location-sharing domain (stderr=0.47). Using both viable domains from the experimental study (social and shopping) results in the same standard error of 0.46. The social media domain performs slighty worse, resulting in a stderr of 0.49. Also here, the SEM with a value of 0.5 is notably higher than most of the regression models Lastly, in the intelligent shopping domain, all single domains produce similar results. Best is the mobile phone domain with a standard error of 0.46 followed by the social media domain; the location-sharing domain is only slightly worse (stderr=0.48). Using all three domains together allows us to reduce the standard error to 0.46. The SEM of this last domain equals to 0.49.

6.1.2 Context-factor-based generic regression analysis (CGR)

Also for the CGR, we used only the context factor instances that were found to be useful in the exploratory study, together with the mean domain privacy levels of the respective domain(s). The results can be found in Table 7. Note that for some domains (for example, the mobile domain for predicting location sharing), none of the context factor instances was found to be suitable in the exporatory study. Those domains are marked using "-" in the table. For a better comparison, we compare the MSE or AUC of the machine learning analysis and the adjusted R^2 for both the MGR and the CGR method in Table 8.

In the location-sharing domain, using all viable context factor coefficients leads to the smallest standard error (stderr=1.21) for this domain, followed by the "topic" or both context factors of the social media domain (stderr=1.22). However, using the "recipient" context factors leads to a higher standard error (stderr=1.27). For the social media domain, equally good results can be achieved using only the location-sharing coefficients (stderr=0.51) or a combination of all viable coefficients (stderr=0.50). If only data from the mobile app domain is available, a prediction with a standard error of 0.56 can be achieved. Data

Table 7 Validation study results for the CGR analysis

Target domain	Coefficients	MSE	AUC	F1	R^2	$adj.R^2$	stderr	F	Sig.
Location	All	1.42	–	–	0.40	0.36	1.21	10.13	< 0.001
	Social - all	1.48	–	–	0.28	0.27	**1.23**	16.89	< 0.001
	Social - topic	1.52	–	–	0.28	0.26	**1.23**	12.65	< 0.001
	Social - recipients	1.82	–	–	0.23	0.21	1.27	18.97	< 0.001
	Mobile - category	–	–	–	–	-			
	Shopping - data type	1.90	–	–	0.15	0.13	1.33	11.29	0.001
	Shopping - stakeholder	–	–	–	–	–			
Social	All	0.30	–	–	0.28	0.25	0.50	9.76	< 0.001
	Mobile - category	0.33	–	–	0.08	0.07	0.56	5.89	0.009
	Shopping - data type	–	–	–	–	–			
	Shopping - stakeholder	–	–	–	–	–			
	Location - all	0.31	–	–	0.23	0.21	**0.51**	12.62	< 0.001
	Location - occasion	0.32	–	–	0.23	0.21	**0.52**	9.40	< 0.001
	Location - requestor	0.32	–	–	0.25	0.22	**0.51**	8.38	< 0.001
Mobile	All	–	0.74	0.59	0.21	0.15	0.46	3.57	0.001
	Social - all	–	0.45	0.13	0.95	0.04	0.49	1.09	0.369
	Social - topic	–	0.52	0.22	0.05	0.02	0.49	1.58	0.182
	Social - recipients	–	0.44	0.13	0.06	0.05	0.49	1.21	0.243
	Location - occasion	–	–	–	–	–			
	Location-requestor	–	0.65	0.45	0.13	0.11	0.47	6.19	< 0.001
	Shopping - all	–	0.66	0.48	0.19	0.15	**0.46**	5.63	< 0.001
	Shopping - data type	–	0.65	0.48	0.19	0.15	**0.46**	5.62	< 0.001
	Shopping - stakeholder	–	0.64	0.52	0.16	0.15	**0.46**	11.81	< 0.001
Shopping	All	–	0.68	0.50	0.18	0.11	0.46	2.6	0.007
	Social - topic	–	–	–	–	-			
	Social - recipients	–	0.60	0.40	0.16	0.15	0.48	2.83	< 0.019
	Location - occasion	–	–	–	–	–			
	Location - requestor	–	0.61	0.42	0.08	0.06	0.47	3.58	0.015
	Mobile - category	–	0.68	0.47	0.10	0.09	**0.46**	7.4	< 0.001

The best prediction source domains are print in bold. The method is only slightly (3%) better in most cases. The results are notably better for the location-sharing domain

from an intelligent retail store did not provide any significant coefficient in the exploratory study and therefore cannot be more precise than a prediction using the MGR approach. For the mobile app domain, using the context factor privacy levels of one or both context factors results in a stderr of 0.46, equal to that of the MGR method. The "requestor" context of the location-sharing domain also leads to good results (stderr=0.47), whereas the other inputs are barely better than random. Similarly, the mobile phone privacy levels lead to best prediction results (stderr=0.46), quickly followed by the "requestor" context-based

Table 8 Comparison of the validation study results for the CGR and MGR analysis and the standard error of the means as a baseline for a prediction without any coefficents

Target domain	Coefficients	CGR stderr	MGR stderr	CGR $adj.R^2$	MGR $adj.R^2$	SEM
Location	All	**1.21**	1.21	0.36	0.30	1.44
	Social - all	1.23	1.27	0.27		
	Social - topic	**1.23**		0.26	0.22	
	Social - recipients	1.27		0.21		
	Mobile - category	–	1.34	–	0.14	
	Shopping - data type	**1.33**	1.34	0.13	0.14	
	Shopping - stakeholder	–		–		
Social	All	0.50	**0.50**	0.25	0.26	0.58
	Mobile - category	0.56	0.56	0.07	0.06	
	Shopping - data type	–	0.54	–	0.13	
	Shopping - stakeholder	–		–		
	Location - all	0.51		0.21		
	Location - occasion	0.52	**0.51**	0.21	0.22	
	Location - requestor	0.51		0.22		
Mobile	All	**0.46**	**0.46**	0.15	0.15	0.5
	Social - all	0.49		0.04		
	Social - topic	0.49	0.49	0.02	0.04	
	Social - recipients	0.49		0.05		
	Location - occasion	–	0.49	–	0.11	
	Location - requestor	**0.47**		0.11		
	Shopping - all	**0.46**	0.46	0.15	0.13	
	Shopping - data type	0.46		0.15		
	Shopping - stakeholder	0.46		0.15		
Shopping	All	**0.46**	**0.46**	0.11	0.13	0.49
	Social - topic	–	**0.47**	–	0.08	
	Social - recipients	0.48		0.15		
	Location - occasion	–	0.48	–	0.05	
	Location - requestor	**0.47**		0.06		
	Mobile - category	**0.46**	**0.46**	0.09	0.10	

privacy levels and the "recipient" privacy levels of the social domain. Comparing the results of the MGR and CGR method, one can see that except for the location domain, both approaches lead to similar results regarding the prediction precision.

6.1.3 Mean-based context-aware regression analysis (MCR)

In contrast to the MGR and CGR method, which has the goal to predict *mean domain privacy levels*, the MCR and CCR methods do a regression on the fine-grained *context-factor-based privacy levels*. We followed the same apporach as

Table 9 Validation study results for the statistical MCR analysis

Target domain	Coefficients	stderr			R^2			$adj.R^2$		
		avg	min	max	avg	min	max	avg	min	max
Location - occasion	All	1.50	1.24	1.75	0.24	0.18	0.28	0.22	0.16	0.26
	Social	**1.55**	1.29	1.82	**0.17**	0.09	0.21	**0.16**	0.09	0.21
	Mobile	1.61	1.33	1.87	0.11	0.03	0.15	0.10	0.02	0.14
	Shopping	1.61	1.35	1.89	0.11	0.07	0.16	0.10	0.06	0.15
Location - requestor	All	1.46	1.40	1.53	0.24	0.19	0.33	0.22	0.17	0.31
	Social	**1.51**	1.45	1.58	**0.17**	0.13	0.23	**0.16**	0.12	0.23
	Mobile	1.57	1.51	1.61	0.11	0.08	0.17	0.10	0.07	0.17
	Shopping	1.57	1.51	1.62	0.11	0.08	0.13	0.10	0.08	0.13
Social - topic	All	0.78	0.65	1.09	0.17	0.05	0.27	0.15	0.02	0.26
	Llocation	**0.79**	0.67	1.10	**0.14**	0.01	0.26	**0.13**	0.00	0.26
	Mobile	0.83	0.70	1.14	0.04	0.00	0.09	0.04	– 0.00	0.08
	Shopping	0.81	0.68	1.11	0.09	0.03	0.14	0.08	0.02	0.13
Social - recipients	All	0.67	0.58	0.79	0.17	0.13	0.24	0.16	0.11	0.23
	Location	**0.68**	0.59	0.81	**0.14**	0.11	0.19	**0.13**	0.10	0.18
	Mobile	0.71	0.61	0.86	0.05	0.02	0.11	0.04	0.01	0.10
	Shopping	0.70	0.62	0.81	0.09	0.06	0.14	0.08	0.05	0.13
Mobile - category	All	0.46	0.43	0.49	0.11	0.06	0.21	0.09	0.04	0.19
	Location	0.46	0.43	0.50	0.07	0.01	0.17	0.06	0.00	0.16
	Social	0.47	0.44	0.50	0.03	0.00	0.06	0.02	– 0.01	0.05
	Shopping	**0.46**	0.44	0.49	**0.06**	0.02	0.10	**0.05**	0.02	0.10
Shopping - data type	All	0.45	0.39	0.50	0.12	0.02	0.22	0.10	– 0.00	0.20
	Location	0.46	0.41	0.50	0.06	0.01	0.11	0.06	0.00	0.11
	Social	0.46	0.41	0.50	0.06	0.01	0.11	0.05	0.00	0.10
	Mobile	**0.46**	0.40	0.50	**0.08**	0.00	0.16	**0.08**	– 0.00	0.16
Shopping - stakeholder	All	0.45	0.42	0.47	0.14	0.04	0.19	0.12	0.02	0.17
	Location	0.46	0.43	0.48	0.07	0.03	0.12	0.06	0.02	0.11
	Social	0.46	0.44	0.48	0.05	0.00	0.07	0.04	– 0.00	0.06
	Mobile	**0.45**	0.43	0.47	**0.10**	0.03	0.13	**0.09**	0.02	0.12

Also here, the prediction is significantly better than random

in the exploratory study, but this time using only the mean domain privacy levels that were found viable in the exploratory study. As the number of context-based privacy levels is very high, we report only the mean, as well as the minimum and maximum of the standard error, R^2 and adjusted R^2 for every domain and context factor for the MCR and CCR method. The results can be seen in Table 9 and 10. The detailed statistical results can be found in Tables 19 (location), 21 (social), 23 (mobile) and 24 (shopping) in the "Appendix". The results of the machine learning analysis are reported in the same short style in Table 12, the

Table 10 Validation study results for the machine learning analysis of the MCR method

Target domain	Coefficients	MSE			AUC			F1		
		avg	min	max	avg	min	max	avg	min	max
Location - occasion	All	2.66	1.80	3.69	–	–	–	–	–	–
	Social	**2.81**	1.90	3.96	–	–	–	–	–	–
	Mobile	2.83	1.90	3.74	–	–	–	–	–	–
	Shopping	2.87	2.12	4.08	–	–	–	–	–	–
Location - requestor	All	2.44	2.21	2.71	–	–	–	–	–	–
	Social	**2.58**	2.37	2.84	–	–	–	–	–	–
	Mobile	2.67	2.31	2.86	–	–	–	–	–	–
	Shopping	2.71	2.50	2.99	–	–	–	–	–	–
Social - topic	All	0.71	0.49	1.25	–	–	–	–	–	–
	Location	**0.69**	0.48	1.34	–	–	–	–	–	–
	Mobile	0.76	0.52	1.46	–	–	–	–	–	–
	Shopping	0.74	0.49	1.27	–	–	–	–	–	–
Social - recipients	All	0.54	0.43	0.73	–	–	–	–	–	–
	Llocation	**0.49**	0.38	0.70	–	–	–	–	–	–
	Mobile	0.56	0.41	0.80	–	–	–	–	–	–
	Shopping	0.54	0.41	0.70	–	–	–	–	–	–
Mobile - category	All	–	–	–	0.66	0.60	0.76	0.34	0.07	0.58
	Location	–	–	–	0.61	0.55	0.70	0.31	0.00	0.64
	Social	–	–	–	0.55	0.35	0.65	0.28	0.00	0.67
	Shopping	–	–	–	**0.62**	0.57	0.68	**0.31**	0.13	0.56
Shopping - data type	All	–	–	–	0.65	0.55	0.77	0.40	0.05	0.81
	Locatiol	–	–	–	0.58	0.43	0.71	0.37	0.00	0.81
	Social	–	–	–	0.57	0.42	0.67	0.28	0.00	0.81
	Mobile	–	–	–	**0.62**	0.48	0.74	**0.40**	0.05	0.81
Shopping - stakeholder	All	–	–	–	0.63	0.52	0.70	0.56	0.47	0.79
	Location	–	–	–	0.62	0.53	0.70	0.43	0.06	0.78
	Social	–	–	–	0.59	0.50	0.66	0.46	0.26	0.80
	Mobile	–	-	–	**0.61**	0.52	0.69	**0.46**	0.15	0.80

complete machine learning results are shown in Tables 20 (location), 22 (social), 23 (mobile) and 25 (shopping) in the "Appendix".

To get a feeling of the quality of the results, we also calculated the precision of a within-domain method for this technique, i.e., a regression and machine learning analysis using *mean domain privacy level* of a domain to predict the context-based privacy levels of the same domain. The standard errors of this method are shown together with the results of the CCR method in Table 13. Detailed results can be found in Tables 15 and 16 in the "Appendix".

For the location-sharing domain, the standard error is similar for both the occasion and requestor context factor instances, whereas the "requestor" privacy levels can be predicted slightly better using all other mean domain privacy levels compared

Table 11 Validation study results for the statistical CCR analysis

Target domain	Coefficients	stderr			R^2			$adj.R^2$		
		avg	min	max	avg	min	max	avg	min	max
Location - occasion	All	1.45	1.18	1.70	0.32	0.21	0.39	0.27	0.15	0.35
	Social	**1.49**	1.25	1.69	**0.22**	0.11	0.30	**0.20**	0.09	0.29
	Mobile	1.69	1.44	1.87	0.08	0.03	0.12	0.07	0.02	0.10
	Shopping	1.63	1.35	1.89	0.13	0.09	0.16	0.11	0.07	0.15
Location - requestor	All	1.39	1.33	1.44	0.34	0.22	0.42	0.30	0.18	0.39
	Social	**1.45**	1.38	1.52	**0.25**	0.16	0.33	**0.23**	0.15	0.29
	Shopping	1.54	1.51	1.58	0.16	0.09	0.23	0.14	0.07	0.20
	Mobile	1.58	1.54	1.61	0.12	0.09	0.17	0.11	0.08	0.16
Social - topic	All	0.76	0.66	0.93	0.25	0.09	0.38	0.20	0.05	0.33
	Mobile	0.79	0.70	0.93	0.06	0.03	0.09	0.05	0.01	0.07
	Shopping	0.78	0.68	0.92	0.09	0.03	0.14	0.07	0.01	0.13
	Location	**0.74**	0.66	0.93	**0.24**	0.17	0.35	**0.22**	0.15	0.33
Social - recipients	All	0.63	0.58	0.71	0.23	0.14	0.37	0.19	0.11	0.33
	Location	**0.65**	0.59	0.79	**0.24**	0.16	0.31	**0.22**	0.14	0.30
	Shopping	0.69	0.61	0.82	0.12	0.09	0.14	0.10	0.07	0.13
	Mobile	0.73	0.67	0.85	0.05	0.04	0.07	0.03	0.02	0.05
Mobile - category	All	0.46	0.43	0.50	0.15	0.07	0.22	0.09	0.00	0.18
	Location	0.46	0.43	0.50	0.09	0.01	0.17	0.07	− 0.00	0.15
	Social	0.47	0.44	0.51	0.04	0.00	0.07	0.02	− 0.02	0.05
	Shopping	**0.47**	0.44	0.50	**0.08**	0.04	0.12	**0.05**	− 0.00	0.10
Shopping - data type	All	0.45	0.38	0.50	0.15	0.03	0.24	0.10	− 0.02	0.22
	Location	0.46	0.41	0.49	0.08	0.03	0.12	0.06	0.01	0.11
	Social	0.47	0.41	0.49	0.07	0.03	0.11	0.05	0.02	0.08
	Mobile	**0.45**	0.39	0.50	**0.10**	0.01	0.19	**0.08**	− 0.02	0.18
Shopping - stakeholder	All	0.45	0.42	0.47	0.16	0.07	0.21	0.11	0.01	0.17
	Location	0.46	0.43	0.48	0.07	0.03	0.13	0.04	− 0.00	0.12
	Social	0.46	0.45	0.48	0.07	0.06	0.07	0.03	0.02	0.04
	Mobile	**0.45**	0.43	0.47	**0.11**	0.05	0.14	**0.08**	0.01	0.12

The method is 2.6–4.4% more precise than its counterpart without context-based privacy levels (MCR)

to the "occasion" context factor. When using only single mean domain privacy levels, the ones from the cluster partner (social media domain; see Sect. 5.1.2) lead to the smallest standard errors. The same holds for the social media domain, where the best domain for predicting the context-based privacy levels is the location-sharing domain. Adding the other mean domain privacy levels does not increase the precision. However, in contrast to the prediction for the location-sharing domain, the standard errors using the shopping *mean domain privacy level* are only slightly higher. For the privacy levels for the different app categories in the mobile app domain, all domains lead to the same precision. Using all of them together again slightly reduces the standard error. Finally, the shopping domain can be predicted

Table 12 Validation study results for the machine learning analysis of the CCR method

Target domain	Coefficients	MSE			AUC			F1		
		avg	min	max	avg	min	max	avg	min	max
Location - occasion	All	2.50	1.72	3.19	–	–	–	–	–	–
	Social	**2.53**	1.75	3.60	–	–	–	–	–	–
	Mobile	3.11	2.19	3.88	–	–	–	–	–	–
	Shopping	2.94	2.09	3.97	–	–	–	–	–	–
Location - requestor	All	2.23	1.93	2.51	–	–	–	–	–	–
	Social	**2.25**	1.99	2.55	–	–	–	–	–	-
	Shopping	2.68	2.40	2.83	–	–	–	–	–	–
	Mobile	2.67	2.34	2.82	–	–	–	–	–	–
Social - topic	All	0.69	0.50	1.07	–	–	–	–	–	–
	Mobile	0.70	0.54	0.97	–	–	–	–	–	–
	Shopping	0.69	0.50	0.94	–	–	–	–	–	–
	Location	**0.63**	0.49	1.02	–	–	–	–	–	–
Social - recipients	All	0.48	0.41	0.55	–	–	–	–	–	-
	Location	**0.49**	0.41	0.71	–	–	–	–	–	–
	Shopping	0.55	0.40	0.73	–	–	–	–	–	–
	Mobile	0.58	0.44	0.78	–	–	–	–	–	–
Mobile - category	All	–	–	–	0.64	0.53	0.76	0.33	0.04	0.66
	Location	–	–	–	0.58	0.47	0.68	0.26	0.00	0.64
	Social	–	–	–	0.56	0.34	0.66	0.20	0.00	0.65
	Shopping	–	–	–	**0.58**	0.52	0.67	**0.29**	0.00	0.48
Shopping - data type	All	–	–	–	0.64	0.48	0.80	0.43	0.00	0.79
	Location	–	–	–	0.60	0.50	0.66	0.25	0.00	0.63
	Social	–	–	–	0.55	0.46	0.68	0.15	0.00	0.50
	Mobile	–	–	–	**0.62**	0.54	0.76	**0.37**	0.04	0.81
Shopping - stakeholder	All	–	–	–	0.56	0.44	0.64	0.48	0.23	0.77
	Location	–	–	–	0.54	0.51	0.60	0.37	0.03	0.80
	Social	–	–	–	0.60	0.52	0.69	0.15	0.14	0.16
	Mobile	–	–	–	**0.61**	0.54	0.67	**0.43**	0.03	0.77

best by both the social media and mobile phone mean domain privacy levels. Using all mean domain privacy levels does not decrease the standard error – neither for the "data type", nor for the "stakeholder" context factor.

Having a look at the results of the within-domain analysis, we can clearly see that the within-domain prediction is notably better than a cross-domain prediction, supporting the findings of earlier research that recommend using cross-domain recommender systems only if no or only sparse within-domain data is available (Adomavicius and Tuzhilin 2005).

Table 13 Comparison of the validation study results for the MCR and CCR analysis

Target domain	Coefficients	MCR			CCR			SEM
		stderr	$adj.R^2$	stderr within -domain	stderr	$adj.R^2$	stderr within -domain	
Location - occasion	All	1.50	0.22	0.90	**1.45**	0.27	0.83	1.70
	Mobile	1.55	0.16		**1.49**	0.20		
	Social	1.61	0.10		1.69	0.07		
	Shopping	1.61	0.10		1.63	0.11		
Location - requestor	All	1.46	0.22	0.81	**1.39**	0.30	0.76	1.66
	Mobile	1.51	0.16		**1.45**	0.23		
	Social	1.57	0.10		1.54	0.14		
	Shopping	1.57	0.10		1.58	0.11		
Social - topic	All	0.78	0.15	0.57	0.76	0.20	0.54	0.84
	Mobile	0.79	0.13		0.79	0.05		
	Location	0.83	0.04		0.78	0.07		
	Shopping	0.81	0.08		**0.74**	0.22		
Social - recipients	All	0.67	0.16	0.44	**0.63**	0.19	0.37	0.73
	Mobile	0.68	0.13		**0.65**	0.22		
	Location	0.71	0.04		0.69	0.10		
	Shopping	0.70	0.08		0.73	0.03		
Mobile - category	All	0.46	0.09	0.40	0.46	0.09	0.39	0.48
	Social	**0.46**	0.06		**0.46**	0.07		
	Location	0.47	0.02		0.47	0.02		
	Shopping	0.46	0.05		0.47	0.05		
Shopping - data type	All	**0.45**	0.10	0.37	**0.45**	0.10	0.36	0.48
	Social	0.46	0.06		0.46	0.06		
	Location	0.46	0.05		0.47	0.05		
	Mobile	0.46	0.08		**0.45**	0.08		
Shopping - stakeholder	All	0.45	0.12	0.37	0.45	0.11	0.32	0.48
	Social	0.46	0.06		0.46	0.04		
	Location	0.46	0.04		0.46	0.03		
	Mobile	**0.45**	0.09		**0.45**	0.08		

6.1.4 Context-factor-based context-aware regression analysis (CCR)

As described in the results section of the exploratory study, we use the context-based privacy levels with $p < 0.1$ for the validation study of the CCR analysis. The standard errors and coefficients of determination of the statistical regression analysis for the different domains are shown in Table 11. Similar to the MCR method, we calculated the average, minimum, and maximum standard error for the different context-factor-based privacy levels, using either the coefficients from all other domains, or only from one of the three other domains. The detailed results for the statistical

and machine learning analysis are shown in Tables 26 (location), 27 (social), 28 (mobile) and 29 (shopping) in the "Appendix". The compact version of the results of the machine learning analysis can be found in Table 12.

A comparison of the prediction precision of the MCR and CCR methods can be found in Table 13. Similar to the MCR method, we performed a statistical and machine learning analysis using the within-domain data, i.e., predicting one context-based privacy level of a domain by using the remaining context-based privacy levels of the same domain. The standard errors can be again found in Table 13, detailed results in Tables 17 and 18 in the "Appendix". Also here, the results show that a within-domain recommendation clearly outperforms a cross-domain recommendation.

For the "occasion" privacy levels in the *location-sharing* domain, the CCR method produces better results for all input domains, except for the mobile app domain. Especially when using all viable coefficients, the standard error can be reduced to 1.50, which is 0.07 better compared to the MCR method. Using only social media coefficients allows us to reduce the stderr by 0.06. Using the context-based privacy levels from the intelligent shopping domain can only improve the stderr by 0.01 compared to the MCR method. Finally, the stderr using the coefficients from the mobile app domain increases by 0.02 to 1.67, which is the highest standard error of all input combinations for the location-sharing domain. A prediction without any input coefficients (i.e., the standard error of the means) ranges from 1.42 to 1.99 (mean 1.70).

The CCR produces better results for the "requestor" context factor of the location-sharing domain as well: When using only social media input, the stderr decreases to 1.48 (−0.06); it stays almost the same for the mobile app (−0.01) and intelligent shopping coefficients (+ − 0.00). Again using all coefficients from all three domains leads to a prediction which is −0.04 more precise compared to the MCR method. The standard error of the means ranges from 1.57 to 1.71 (average 1.66).

The social media domain can also profit from the more fine-grained input of the CCR domain: Using all coefficients, we can reduce the stderr by 0.04 for the "topic", and by 0.05 for the "recipients" context factor. Also using only the coefficients from the location-sharing domain reduces the standard errors for both "topic" (−0.03) and "recipients" (−0.02). However, the prediction using only mobile app (− 0.01 for "topic", + − 0.00 for "recipients") or intelligent shopping (+ − 0.00 for "topic", − 0.01 for "recipients") coefficients improves only slightly. The standard error of the means ranges from 0.64 to 0.87 (mean 0.73) for the "recipients" and from 0.72 to 1.13 (mean 0.84) for the "topic" context factor.

In the intelligent shopping domain, only the precision using all context-based privacy levels improved the regression precision (− 0.03 for "stakeholder", − 0.02 for "data type"). The precision using only coefficients from either the social media (+-0.00 for "stakeholder", −0.01 for "data type"), location sharing (+ − 0.00 for "stakeholder", + − 0.00 for "data type") or mobile app (+0.01 for "stakeholder", − 0.01 for "data type") domain did not change much.

The standard error of the means ranges from 0.42 to 0.5 (mean 0.48) for the datatypes, and from 0.46 to 0.49 (mean 0.48) for the stakeholders.

Finally, the mobile app domain also could not profit significantly from the increased granularity of the CCR method. Using coefficients only from the social media (0.01) or location-sharing (− 0.01) domain only slightly decreases the standard error, whereas it stays the same for the intelligent shopping domain and also when using all domains together as an input. The standard error of the means of this domain ranges from 0.45 to 0.50 (mean 0.48).

7 Discussion

7.1 Predicting mean domain privacy levels using MGR VS CGR

We presented two different approaches for predicting the *mean domain privacy level* (the mean privacy level computed over all privacy levels of a domain) of a domain: first the MGR approach that uses the other mean domain privacy levels as an input, and second the CGR method that uses the privacy levels for the different context factor instances (like the privacy level given for social network posts about food, or a location-sharing privacy level that has to be applied when a family member requests the location) in addition to the mean-based privacy level. Usually, one would think that more data leads to a higher precision (e.g., a lower standard error). However, this is only the case for the location-sharing domain, where the CGR method leads to lower standard errors when using the social media or shopping privacy levels, and especially when all domain data can be used. Still, the size of the effect is relatively small, with a standard error improvement ranging up to 0.04 (or 3%) when using the social media privacy levels as an input, resulting in a final standard error of 1.23. For all other domains, the CGR method offers no improvement compared to the simpler MGR approach.

Having a look at the coefficients used for the location-sharing prediction (Table 4), we can see that there are two context factor instances which are of major importance from the social media domain: the "topic" context factor instances "movies" and especially "events". Those two post topics are occasions in which users typically also share their location (especially on events), which might lead to their suitability for a prediction, which leads to a decreased standard error when added to the set of coefficients. The same seems to hold for the amount of items bought, which is used for predicting the location-sharing level when only shopping privacy levels are available. People like to share their location during shopping either when they have bought expensive products, or when they have bought an extraordinarily high amount of items, for example, at a sale or at a factory outlet store. To conclude, we can state that the simple MGR approach works very well for most of the domains. There are only some domains where the increased data set of the CGR method can improve the precision, like the location-sharing domain. Which other domains are also suitable for CGR should be a research topic of future work. As a rule of thumb, it seems like CGR can profit from its context-based privacy levels from the other

domains, if some of them are very similar or are often used together with the privacy levels of the target domain.

7.2 Predicting context-based privacy levels using MCR VS CCR

When it comes to predicting the context-based privacy levels, e.g., the different privacy levels depending on the context factors mentioned in Sect. 3.1, the simplistic mean domain privacy level-based approach (MCR) still performs well. However, in this case, the context-based method (CCR) can outperform the MCR in most cases, supporting earlier work stating that privacy decisions are multidimensional in general and should take context factors into account, rather than predicting a general privacy decision (Knijnenburg et al. 2013). In the location-sharing domain, where the CGR already outperformed the MGR, the CCR leads to a standard error that is on average 2.6% – 4.4% more precise than the MCR approach when using all other domains as an input. The improvement is even larger for the social media domain, where the standard error is reduced by 5% – 7% when using all input data. For the shopping domain, the improvement amounts to 3.2% – 4.3%. The mobile app category remains the only one where the CCR approach led to only meaningless improvements. In the domain of mobile app permissions, the context-based privacy levels were based on the app category as a context factor. However, none of the other domains have a similar context factor, so the other context-based privacy levels, which were based on the requestor or the occasion, for example, were not of any help, so that the CGR approach could not lead to an improvement of the prediction precision. We therefore speculate that the performance of the CCR depends on the semantic distance between the context factors of the target domain and the input domains. Semantic similarity in this context means that the context factor values often have the same meaning, e.g., the occasions when a location is shared are often similar to topics of a social network post. A social network post, for example, can be about a new movie the user just watched at home (topic "movies"); but a location can also be shared when going to the cinema with friends (occasion "movies"), forming a semantic similarity between the two context factors. Heckmann's work also indicates that the decision which information to show to the user depends on similar context, and offers a reasoner to find the context that is most semantically similar (Heckmann 2006).

Whether this assumption can be generalized, should be checked in future work. In general, we conclude that, for predicting context-based privacy settings, the context-based CCR method should be preferred for most cases, if the data is available. But if the context factors do not match well, e.g., the semantic distance between them is high, the CCR seems not to lead to any advantage. However, the simplistic MCR approach that uses only the mean domain privacy levels from the other domains, performs surprisingly well, even for predicting fine-grained privacy levels.

7.3 Size of the effect and user acceptance

Both the MGR and CGR led to a prediction precision which is better than the standard error of the means, i.e., a prediction without coefficients, showing that the usage of privacy levels from other domains allows a prediction significantly better than random. Whether the achieved precision is sufficient to be accepted by users still has to be checked in future studies.

Also with the MCR and CCR methods, the prediction precision was notably better for the location domain compared to the standard error of the means. For the other social media domain, the difference between the prediction precision is only slightly (about 5 to 10%) better, whereas no improvement can be achieved for the mobile phone and shopping domain compared to the MCR approach. All approaches are about 15% to 20% more precise than an unparameterized, general prediction using no input. However, the within-domain predictions, which form an upper bound for the prediction precision of the cross-domain recommenders, can also decrease the standard error only by about 20 to 25% in most cases. Given that fact, the improvement of the standard error can be seen as normal for a cross-domain recommender, which is typically performing significantly worse than a within-domain recommender system. Nevertheless, whether such a recommender system would be accepted by users, and also whether a within-domain recommender would lead to results that are perceived as better or more useful by the users, has to be investigated in future work.

7.4 Which data set is to be used for a prediction

After deciding on a suitable predicting method (either mean-based or context-based), the next question is which data should be used for the prediction, or whether the available data is sufficient for a prediction. In general, if data from all other domains is available, this data should also be used. In our experiments, we did not identify any case where the precision decreased when using all domain data instead of only a specific domain. If only privacy levels of some of the domains are available, or if the data has to be acquired/processed first, it is best to think in clusters. In the experiments, we identified two domain clusters, inside which each domain is particularly suitable for predicting the other domain. The first is the location-sharing/social media cluster, the second is the mobile app permission/intelligent shopping cluster. Whenever the privacy levels from the cluster partner are available or can be acquired, they should be preferred before those of all other domains. If data from other domains is already available, it should be added as well, although it will not increase the precision very much. We recommend not to add further domain data if the additional data must be acquired first, and the acquisition would lead to an increased user burden or an immense computing overhead.

Table 14 Summary of the key findings of the discussion

MGR VS CGR	- MGR and CGR similar in most domains - Precision about 10%-15% better than random - CGR usually only slightly better than MGR(3%) - CGR notably better than MGR in location-sharing domain - Notably better than standard error of the means, i.e., better than prediction without input - Whether precision is enough to be accepted by users has to be investigated in future work
MCR VS CCR	- MCR still performs well CCR on average about 5%-10% better than MCR - CCR outperforms MCR in location and shopping domain, could be caused by semantic similarity - Notably better than standard error of the means, i.e., better than prediction without input - Whether precision is enough to be accepted by users has to be investigated in future work
best input data	Two clusters exist, that allow to predict each other: location sharing/social media and mobile phone/shopping
privacy paradox of recommenders	- Additional private data needed for recommending privacy settings - Techniques exist that obfuscate private data whitout disturbing recommender system (e.g., differential privacy, PLA-based systems)

7.5 The privacy paradox in privacy recommender systems

The goal of our research is to help users to tune their privacy settings so that they disclose as little private information as needed while still keeping the services (for example, the social network or smartphone) usable. However, in order to allow an automatic prediction of privacy settings, we in fact *need additional information from the user* as an input for a meaningful prediction. This fact that we call the "recommender systems privacy paradox" has been part of research since several years (Toch et al. 2012). There are several approaches that allow the user to increase her privacy in such a recommender system, by aggregating the personal information together with 2, 3, 4 or n other data sets at the cost of prediction precision (Toch et al. 2012). Data expiration and data morphing are further methods that can enhance privacy at the cost of the recommendation quality (Toch et al. 2012). Other approaches try to adapt the recommender system itself to be more privacy-aware, for example, by employing a differential privacy mechanism in matrix factorisation approaches (Friedman et al. 2016). Another PLA-based framework selects a personalisation method at runtime that fits the user's privacy requirements, to enhance privacy while keeping the prediction quality at a similar level (Wang and Kobsa 2013). In our studies, we were focused on finding out whether and how good privacy recommendations work in an *optimal* case, where the user's personal information is fully available. However, we like to check how well our approaches work when privacy-enhancing techniques are included in future work. They key findings of the discussion are summarized in Table 14.

7.6 Future work

We have conducted an experimental study to find out which domains and which privacy levels are viable inputs for a prediction. Both the experimental and the validation study have been conducted with a data set of sufficient size for a regression analysis. However, related work has shown that the prediction precision can be further increased involving a large data set, containing thousands or even millions of users (Liu et al. 2016). We therefore would like to extend the experiment in the future to larger data sets, so we can investigate to which extent the prediction precision can be increased.

We found differences in the prediction precision between the methods using the *mean domain privacy level* as an input (MGR and MCR) and the methods using context-based privacy levels (CGR and CCR). In some domains, the CGR and the CCR outperform the MGR and MCR. However, the difference is below 10% and therefore relatively small. In future work, we would therefore like to perform a field study, where users have to use a social media account, for example, including privacy settings based on one of the aforementioned approaches, and the task to use the account for some weeks and to adapt the privacy settings, if needed. At the end of the study, we will compare the changes to the privacy settings and thereby the number of errors made by each method. Using a questionnaire, we will evaluate the subjective differences on the perceived prediction precision.

We were able to identify several coefficients for the context-based regression methods (MCR and CCR) that can be of use in future work. We found that the four domains treated in this article can be clustered into two clusters, inside which a regression of the partner's privacy levels is possible. However, there are plenty of other domains that have not been part of our research so far. In future work, we therefore want to investigate whether other domains also form clusters together, or whether they are part of one of the two aforementioned clusters. The ultimate goal is to find out whether there is a finite number of clusters that allow a prediction of each other's privacy levels, what domains they include, and which common properties they share that make them belong to the same cluster. Apart from that, we investigated only a finite number of context factors in this article that have been found to be significant in related work. There might be still other context factors that have not been discovered yet, which we would propose as topics for future work.

8 Conclusion

Users often neglect their privacy settings, as they often do not see the potential risks that come with oversharing of the data. There exist a lot of solutions that try to aid the user in choosing the privacy settings using machine learning, either by using other privacy settings from the same domain as an input, or by utilizing the user's personality and privacy attitudes for a personalized recommendation. However, as this information is not always available, we examined whether privacy settings from other domains can also be used as an input for the prediction. We investigated the prediction of a *mean domain privacy level* that gives only one general user-specific privacy level for a domain as an orientation, as well as the prediction of fine-grained context-based privacy levels that give a distinct personalized privacy level for each combination of context factors. We found all context factors except for the permission type in the mobile phone domain to be significantly different and therefore as potentially suitable input for a prediction. The results show that both types of privacy levels can be predicted already using only the mean domain privacy levels from the other domains. However, the fine-grained context-based privacy levels and the mean domain privacy levels from the location-sharing domain can be predicted better using the context-based privacy levels as an input. Although we verified the selected regression coefficients within a validation study and although we achieved a small increase in the prediction precision using the CGR and CCR method compared to the MGR and MCR method in some cases, we would like to test the suitability of our prediction in future work in an in-the-wild study, where the privacy levels are predicted from actual privacy settings of the users, and check how well the implementation of the predicted privacy levels in the different domains fits the actual desired privacy settings of the user.

Appendix

See Fig. 5.

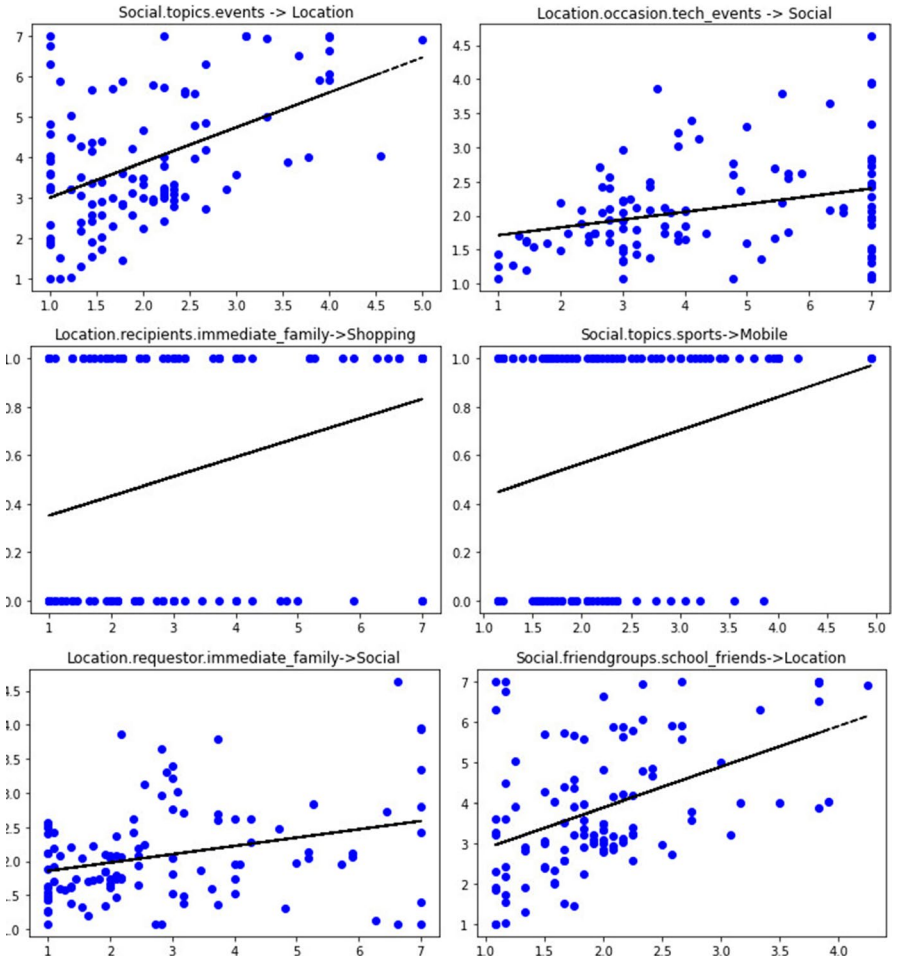

Fig. 5 Selection of scatter plots of the coefficents used for the CGR and CCR analyses

See Tables 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28 and 29.

Table 15 Statistical results of the within-domain analysis for the MCR method

Domain	stderr			R^2			$adj.R^2$		
	Avg.	Min	Max	Avg	Min	Max	Avg	Min	Max
Location - occasion	0.90	0.70	1.18	0.71	0.48	0.81	0.71	0.48	0.81
Location - requestor	0.81	0.63	0.95	0.76	0.65	0.85	0.75	0.65	0.84
Social - topics	0.57	0.35	0.93	0.55	0.30	0.78	0.55	0.30	0.78
Social - recipients	0.44	0.33	0.56	0.63	0.58	0.74	0.63	0.57	0.74
Mobile - category	0.40	0.38	0.43	0.30	0.21	0.36	0.30	0.20	0.36
Shopping - data type	0.37	0.33	0.42	0.38	0.22	0.51	0.37	0.21	0.51
Shopping - stakeholder	0.37	0.33	0.40	0.38	0.29	0.50	0.38	0.28	0.50

Table 17 Statistical results of the within-domain analysis for the CCR method

Domain	stderr			R^2			$adj.R^2$		
	Avg.	Min	Max	Avg.	Min	Max	Avg.	Min	Max
Location - occasion	0.83	0.60	1.06	0.76	0.55	0.84	0.75	0.53	0.83
Location - requestor	0.76	0.64	0.96	0.79	0.68	0.85	0.78	0.67	0.85
Social - topic	0.54	0.35	0.90	0.61	0.40	0.79	0.59	0.37	0.78
Social - recipients	0.37	0.27	0.53	0.74	0.64	0.84	0.74	0.63	0.84
Mobile - category	0.39	0.38	0.40	0.35	0.29	0.39	0.33	0.28	0.38
Shopping - data type	0.36	0.31	0.41	0.45	0.24	0.62	0.43	0.23	0.61
Shopping - stakeholder	0.32	0.27	0.39	0.56	0.38	0.67	0.55	0.37	0.66

Table 16 Machine learning results of the within-domain analysis for the MCR method

Domain	MSE			AUC			F1		
	Avg.	Min	Max	Avg.	Min	Max	Avg.	Min	Max
Location - occasion	0.78	0.42	1.48	–	–	–	–	–	–
Location - requestor	0.69	0.42	0.95	–	–	–	–	–	–
Social - topics	0.42	0.18	0.91	–	–	–	–	–	–
Social - recipients	0.26	0.13	0.37	–	–	–	–	–	–
Mobile - category	–	–	–	0.79	0.72	0.83	0.59	0.31	0.75
Shopping - data type	–	–	–	0.83	0.71	0.91	0.72	0.49	0.82
Shopping - stakeholder	–	–	–	0.84	0.74	0.93	0.75	0.69	0.86

Table 18 Machine learning results of the within-domain analysis for the CCR method

Domain	MSE			AUC			F1		
	avg	min	max	avg	min	max	avg	min	max
Location - occasion	0.93	0.52	1.34	–	–	–	–	–	–
Location - requestor	0.68	0.39	0.95	–	–	–	–	–	–
Social - topic	0.47	0.24	1.05	–	–	–	–	–	–
Social - recipients	0.24	0.12	0.34	–	–	–	–	–	–
Mobile - category	–	–	–	0.82	0.79	0.86	0.56	0.29	0.76
Shopping - data type	–	–	–	0.87	0.76	0.98	0.72	0.55	0.91
Shopping - stakeholder	–	–	–	0.93	0.84	0.98	0.81	0.64	0.89

Table 19 Detailed validation study results for the statistical and machine learning analysis of the MCR method for the location domain

Target domain	Coefficients	stderr	R^2	R^2_{adj}	F	p	MSE
Location_t_famil	All	1.34	0.24	0.22	13.25	< 0.001	2.05
Location_t_famil	Social-all	1.40	0.15	0.15	22.94	< 0.001	2.19
Location_t_famil	Mobile-all	1.43	0.12	0.11	16.55	< 0.001	2.09
Location_t_famil	shopping-all	1.41	0.14	0.14	21.10	< 0.001	2.15
Location_t_event	All	1.24	0.25	0.23	14.03	< 0.001	1.80
Location_t_event	Social-all	1.29	0.18	0.18	28.52	< 0.001	1.90
Location_t_event	Mobile-all	1.33	0.13	0.12	18.33	< 0.001	1.89
Location_t_event	Shopping-all	1.35	0.10	0.09	14.03	< 0.001	2.12
Location_t_movie	All	1.66	0.23	0.22	12.69	< 0.001	3.13
Location_t_movie	Social-all	1.70	0.18	0.17	28.05	< 0.001	3.37
Location_t_movie	Mobile-all	1.77	0.11	0.10	15.23	< 0.001	3.35
Location_t_movie	Shopping-all	1.80	0.08	0.08	11.50	< 0.001	3.65
Location_t_polit	All	1.61	0.19	0.17	9.82	< 0.001	2.98
Location_t_polit	Social-all	1.62	0.17	0.16	25.42	< 0.001	3.01
Location_t_polit	Mobile-all	1.75	0.03	0.02	3.55	0.06	3.34
Location_t_polit	Shopping-all	1.70	0.09	0.08	12.25	< 0.001	2.90
Location_t_food	All	1.52	0.28	0.26	16.11	< 0.001	2.64
Location_t_food	Social-all	1.62	0.17	0.16	25.86	< 0.001	2.92
Location_t_food	Mobile-all	1.64	0.15	0.14	22.16	< 0.001	2.82
Location_t_food	Shopping-all	1.63	0.16	0.15	23.80	< 0.001	2.88
Location_t_work	All	1.43	0.26	0.24	14.36	< 0.001	2.38
Location_t_work	Social-all	1.47	0.20	0.20	32.54	< 0.001	2.56
Location_t_work	Mobile-all	1.56	0.09	0.09	13.23	< 0.001	2.85
Location_t_work	Shopping-all	1.55	0.12	0.11	16.78	< 0.001	2.65
Location_t_hobbi	All	1.44	0.23	0.21	12.42	< 0.001	2.44
Location_t_hobbi	Social-all	1.50	0.15	0.14	21.75	< 0.001	2.59
Location_t_hobbi	Mobile-all	1.52	0.13	0.12	18.26	< 0.001	2.39
Location_t_hobbi	Shopping-all	1.53	0.12	0.11	16.53	< 0.001	2.46
Location_t_music	All	1.74	0.23	0.21	12.63	< 0.001	3.41
Location_t_music	Social-all	1.78	0.18	0.18	28.63	< 0.001	3.66
Location_t_music	Mobile-all	1.87	0.10	0.10	14.52	< 0.001	3.74
Location_t_music	Shopping-all	1.89	0.08	0.07	11.13	0.00	3.82
Location_t_trave	All	1.38	0.18	0.16	8.92	< 0.001	2.29
Location_t_trave	Social-all	1.44	0.09	0.09	12.92	< 0.001	2.33
Location_t_trave	Mobile-all	1.41	0.13	0.12	18.44	< 0.001	2.27
Location_t_trave	Shopping-all	1.46	0.07	0.06	9.03	0.00	2.34
Location_t_sport	All	1.41	0.27	0.25	15.23	< 0.001	2.47
Location_t_sport	Social-all	1.45	0.21	0.21	34.34	< 0.001	2.38
Location_t_sport	Mobile-all	1.55	0.09	0.08	12.68	< 0.001	2.67
Location_t_sport	Shopping-all	1.52	0.13	0.12	19.04	< 0.001	2.57
Location_t_tech	All	1.75	0.24	0.22	13.28	< 0.001	3.69

Table 19 (continued)

Target domain	Coefficients	stderr	R^2	R^2_{adj}	F	p	MSE
Location_t_tech	Social-all	1.82	0.17	0.16	25.60	< 0.001	3.97
Location_t_tech	Mobile-all	1.87	0.12	0.12	17.66	< 0.001	3.70
Location_t_tech	Shopping-all	1.88	0.11	0.10	15.97	< 0.001	4.08

Table 20 Detailed validation study results for the statistical and machine learning analysis of the MCR method for the location domain continued

Target domain	Coefficients	stderr	R^2	R^2_{adj}	F	p	MSE
Location_r_acqu	All	1.51	0.24	0.22	12.94	< 0.001	2.41
Location_r_acqu	Social-all	1.55	0.17	0.17	26.90	< 0.001	2.60
Location_r_acqu	Mobile-all	1.61	0.11	0.11	16.21	< 0.001	2.71
Location_r_acqu	Shopping-all	1.62	0.10	0.09	13.94	< 0.001	2.77
Location_r_school	All	1.44	0.21	0.19	11.09	< 0.001	2.21
Location_r_school	Social-all	1.48	0.15	0.14	22.62	< 0.001	2.45
Location_r_school	Mobile-all	1.53	0.09	0.09	13.20	< 0.001	2.42
Location_r_school	Shopping-all	1.52	0.10	0.10	14.85	< 0.001	2.55
Location_r_online	All	1.43	0.19	0.17	9.67	< 0.001	2.30
Location_r_online	Social-all	1.45	0.15	0.14	21.57	< 0.001	2.37
Location_r_online	Mobile-all	1.51	0.08	0.07	10.89	0.00	2.63
Location_r_online	Shopping-all	1.51	0.08	0.08	11.45	< 0.001	2.50
Location_r_sports	All	1.43	0.24	0.22	13.08	< 0.001	2.41
Location_r_sports	Social-all	1.47	0.19	0.18	29.03	< 0.001	2.51
Location_r_sports	Mobile-all	1.55	0.09	0.08	12.28	< 0.001	2.74
Location_r_sports	Shopping-all	1.53	0.12	0.11	16.67	< 0.001	2.64
Location_r_exfam	All	1.50	0.25	0.23	13.64	< 0.001	2.71
Location_r_exfam	Social-all	1.56	0.17	0.16	26.27	< 0.001	2.64
Location_r_exfam	Mobile-all	1.60	0.13	0.13	19.30	< 0.001	2.72
Location_r_exfam	Shopping-all	1.62	0.10	0.09	14.10	< 0.001	2.71
Location_r_imfam	All	1.53	0.21	0.19	10.90	< 0.001	2.68
Location_r_imfam	Social-all	1.58	0.14	0.13	20.75	< 0.001	2.76
Location_r_imfam	Mobile-all	1.61	0.11	0.10	14.99	< 0.001	2.81
Location_r_imfam	Shopping-all	1.62	0.10	0.09	14.27	< 0.001	2.99
Location_r_work	All	1.43	0.26	0.24	14.36	< 0.001	2.38
Location_r_work	Social-all	1.47	0.20	0.20	32.54	< 0.001	2.56
Location_r_work	Mobile-all	1.56	0.09	0.09	13.23	< 0.001	2.85
Location_r_work	Shopping-all	1.55	0.12	0.11	16.78	< 0.001	2.65
Location_r_closef	All	1.51	0.22	0.20	11.51	< 0.001	2.60
Location_r_closef	Social-all	1.57	0.13	0.12	19.24	< 0.001	2.84
Location_r_closef	Mobile-all	1.60	0.10	0.10	14.63	< 0.001	2.86
Location_r_closef	Shopping-all	1.57	0.13	0.13	19.80	< 0.001	2.73
Location_r_friends	All	1.40	0.33	0.31	20.24	< 0.001	2.23
Location_r_friends	Social-all	1.48	0.23	0.23	38.29	< 0.001	2.48
Location_r_friends	Mobile-all	1.54	0.17	0.17	26.31	< 0.001	2.31
Location_r_friends	Shopping-all	1.58	0.13	0.12	18.73	< 0.001	2.82

Table 21 Detailed validation study results for the statistical and machine learning analysis of the MCR method for the Social media domain

Target domain	Coefficients	stderr	R^2	R^2_{adj}	F	p	MSE
Social_t_famil	All	0.92	0.05	0.02	2.05	0.11	1.00
Social_t_famil	Location-all	0.93	0.01	0.00	1.54	0.22	1.00
Social_t_famil	Mobile-all	0.92	0.03	0.03	4.42	0.04	0.96
Social_t_famil	Shopping-all	0.92	0.03	0.03	4.39	0.04	1.02
Social_t_event	All	0.73	0.21	0.19	11.06	< 0.001	0.66
Social_t_event	Location-all	0.74	0.18	0.17	28.03	< 0.001	0.58
Social_t_event	Mobile-all	0.79	0.07	0.06	9.37	0.00	0.66
Social_t_event	Shopping-all	0.77	0.10	0.09	13.54	< 0.001	0.71
Social_t_movie	All	0.70	0.27	0.26	15.60	< 0.001	0.49
Social_t_movie	Location-all	0.70	0.26	0.26	44.81	< 0.001	0.53
Social_t_movie	Mobile-all	0.78	0.07	0.07	9.90	0.00	0.63
Social_t_movie	Shopping-all	0.78	0.08	0.07	11.05	0.00	0.63
Social_t_polit	All	1.09	0.09	0.07	4.32	0.01	1.25
Social_t_polit	Location-all	1.10	0.06	0.05	8.37	0.00	1.34
Social_t_polit	Mobile-all	1.14	0.00	−0.00	0.58	0.45	1.46
Social_t_polit	Shopping-all	1.11	0.06	0.05	7.74	0.01	1.27
Social_t_food	All	0.68	0.20	0.18	10.65	< 0.001	0.54
Social_t_food	Location-all	0.71	0.13	0.12	19.07	< 0.001	0.55
Social_t_food	Mobile-all	0.72	0.09	0.08	12.19	< 0.001	0.58
Social_t_food	Shopping-all	0.70	0.14	0.13	20.60	< 0.001	0.53
Social_t_work	All	0.76	0.24	0.23	13.48	< 0.001	0.64
Social_t_work	Location-all	0.78	0.19	0.18	29.99	< 0.001	0.60
Social_t_work	Mobile-all	0.86	0.03	0.02	4.18	0.04	0.79
Social_t_work	Shopping-all	0.81	0.14	0.13	20.43	< 0.001	0.70
Social_t_hobbi	All	0.65	0.19	0.17	9.88	< 0.001	0.53
Social_t_hobbi	Location-all	0.67	0.15	0.14	21.91	< 0.001	0.48
Social_t_hobbi	Mobile-all	0.70	0.06	0.05	8.05	0.01	0.52
Social_t_hobbi	Shopping-all	0.68	0.11	0.11	16.47	< 0.001	0.49
Social_t_music	All	0.70	0.13	0.11	6.13	< 0.001	0.66
Social_t_music	Location-all	0.70	0.11	0.10	15.99	< 0.001	0.57
Social_t_music	Mobile-all	0.73	0.03	0.02	3.66	0.06	0.61
Social_t_music	Shopping-all	0.72	0.06	0.05	8.21	0.00	0.58
Social_t_trave	All	0.80	0.08	0.06	3.52	0.02	0.82
Social_t_trave	Location-all	0.80	0.07	0.07	10.14	0.00	0.77
Social_t_trave	Mobile-all	0.82	0.02	0.01	2.30	0.13	0.76
Social_t_trave	Shopping-all	0.82	0.03	0.02	3.42	0.07	0.84
Social_t_sport	All	0.67	0.22	0.20	11.70	< 0.001	0.52
Social_t_sport	Location-all	0.68	0.19	0.18	28.97	< 0.001	0.48
Social_t_sport	Mobile-all	0.73	0.05	0.04	6.61	0.01	0.58
Social_t_sport	Shopping-all	0.71	0.11	0.10	15.99	< 0.001	0.52
Social_t_tech	All	0.84	0.19	0.17	9.82	< 0.001	0.75

Table 21 (continued)

Target domain	Coefficients	stderr	R^2	R^2_{adj}	F	p	MSE
Social_t_tech	Location-all	0.85	0.16	0.16	25.08	< 0.001	0.75
Social_t_tech	Mobile-all	0.91	0.04	0.03	5.22	0.02	0.87
Social_t_tech	Shopping-all	0.88	0.09	0.09	12.91	< 0.001	0.83
Social_r_acqu	All	0.66	0.18	0.16	9.00	< 0.001	0.49
Social_r_acqu	Location-all	0.66	0.15	0.14	22.08	< 0.001	0.45
Social_r_acqu	Mobile-all	0.69	0.06	0.06	8.81	0.00	0.57
Social_r_acqu	Shopping-all	0.69	0.09	0.08	11.94	< 0.001	0.53

Table 22 Detailed validation study results for the statistical and machine learning analysis of the MCR method for the Social media domain continued

Target domain	Coefficients	stderr	R^2	R^2_{adj}	F	p	MSE
Social_r_school	All	0.58	0.21	0.19	10.86	< 0.001	0.43
Social_r_school	Location-all	0.59	0.16	0.15	23.59	< 0.001	0.38
Social_r_school	Mobile-all	0.61	0.11	0.10	15.69	< 0.001	0.41
Social_r_school	Shopping-all	0.62	0.09	0.09	13.23	< 0.001	0.46
Social_r_online	All	0.68	0.14	0.12	6.65	< 0.001	0.50
Social_r_online	Location-all	0.68	0.11	0.10	15.15	< 0.001	0.49
Social_r_online	Mobile-all	0.70	0.05	0.05	7.16	0.01	0.54
Social_r_online	Shopping-all	0.69	0.08	0.07	10.54	0.00	0.52
Social_r_sports	All	0.79	0.17	0.15	8.46	< 0.001	0.73
Social_r_sports	Location-all	0.81	0.12	0.11	17.39	< 0.001	0.70
Social_r_sports	Mobile-all	0.85	0.03	0.02	3.75	0.05	0.80
Social_r_sports	Shopping-all	0.81	0.11	0.10	15.92	< 0.001	0.70
Social_r_exfam	All	0.62	0.16	0.14	8.16	< 0.001	0.45
Social_r_exfam	Location-all	0.62	0.15	0.14	22.59	< 0.001	0.41
Social_r_exfam	Mobile-all	0.67	0.02	0.01	2.83	0.09	0.47
Social_r_exfam	Shopping-all	0.65	0.06	0.05	8.24	0.00	0.47
Social_r_imfam	All	0.58	0.19	0.17	9.56	< 0.001	0.43
Social_r_imfam	Location-all	0.59	0.17	0.16	25.80	< 0.001	0.38
Social_r_imfam	Mobile-all	0.63	0.04	0.04	5.95	0.02	0.43
Social_r_imfam	Shopping-all	0.62	0.08	0.07	10.78	0.00	0.41
Social_r_work	All	0.76	0.24	0.23	13.48	< 0.001	0.64
Social_r_work	Location-all	0.78	0.19	0.18	29.99	< 0.001	0.60
Social_r_work	Mobile-all	0.86	0.03	0.02	4.18	0.04	0.79
Social_r_work	Shopping-all	0.81	0.14	0.13	20.43	< 0.001	0.70
Social_r_closef	All	0.66	0.15	0.13	7.42	< 0.001	0.55
Social_r_closef	Location-all	0.66	0.13	0.12	19.20	< 0.001	0.49
Social_r_closef	Mobile-all	0.69	0.05	0.04	6.63	0.01	0.50
Social_r_closef	Shopping-all	0.68	0.07	0.06	9.24	0.00	0.51
Social_r_friends	All	0.70	0.13	0.11	6.24	< 0.001	0.64
Social_r_friends	Location-all	0.70	0.12	0.11	16.68	< 0.001	0.51
Social_r_friends	Mobile-all	0.73	0.02	0.02	3.09	0.08	0.54
Social_r_friends	Shopping-all	0.72	0.06	0.05	7.74	0.01	0.54

Table 23 Detailed validation study results for the statistical and machine learning analysis of the MCR method for the mobile phone domain

Target domain	Coefficients	stderr	R^2	R^2_{adj}	F	p	MSE
Mobile_c_comm	All	0.45	0.09	0.07	4.10	0.01	–
Mobile_c_comm	Location-all	0.46	0.01	0.00	1.39	0.24	–
Mobile_c_comm	Social-all	0.45	0.05	0.05	7.35	0.01	–
Mobile_c_comm	Shopping-all	0.45	0.06	0.06	8.69	0.00	–
Mobile_c_edu	All	0.43	0.11	0.09	5.26	0.00	–
Mobile_c_edu	Location–all	0.43	0.08	0.07	11.14	0.00	–
Mobile_c_edu	Social-all	0.44	0.06	0.05	8.21	0.00	–
Mobile_c_edu	Shopping-all	0.44	0.06	0.05	8.36	0.00	–
Mobile_c_news	All	0.45	0.21	0.19	11.14	< 0.001	–
Mobile_c_news	Location-all	0.45	0.17	0.16	25.72	< 0.001	–
Mobile_c_news	Social-all	0.49	0.02	0.01	2.60	0.11	–
Mobile_c_news	Shopping-all	0.47	0.10	0.10	14.48	< 0.001	–
Mobile_c_shopp	All	0.48	0.11	0.09	5.00	0.00	–
Mobile_c_shopp	Location-all	0.49	0.05	0.04	6.24	0.01	–
Mobile_c_shopp	Social-all	0.49	0.04	0.04	5.65	0.02	–
Mobile_c_shopp	Shopping-all	0.48	0.09	0.08	12.76	< 0.001	–
Mobile_c_Social	All	0.45	0.14	0.12	6.88	< 0.001	–
Mobile_c_Social	Location-all	0.45	0.13	0.12	19.19	< 0.001	–
Mobile_c_Social	Social-all	0.48	0.02	0.01	2.51	0.12	–
Mobile_c_Social	Shopping-all	0.47	0.05	0.04	6.17	0.01	–
Mobile_c_tools	All	0.45	0.08	0.05	3.41	0.02	–
Mobile_c_tools	Location-all	0.46	0.04	0.03	5.40	0.02	–
Mobile_c_tools	Social–all	0.46	0.03	0.02	3.44	0.07	–
Mobile_c_tools	Shopping-all	0.45	0.06	0.05	8.26	0.00	–
Mobile_c_travel	All	0.44	0.07	0.05	3.18	0.03	–
Mobile_c_travel	Location-all	0.43	0.06	0.05	8.32	0.00	–
Mobile_c_travel	Social-all	0.45	0.01	⁻0.00	0.64	0.43	–
Mobile_c_travel	Shopping-all	0.44	0.02	0.02	3.13	0.08	–
Mobile_c_games	All	0.49	0.06	0.04	2.78	0.04	–
Mobile_c_games	Location-all	0.50	0.03	0.02	4.10	0.05	–
Mobile_c_games	Social-all	0.50	0.00	⁻0.01	0.01	0.93	–
Mobile_c_games	Shopping-all	0.49	0.03	0.03	4.60	0.03	–

Table 24 Detailed validation study results for the statistical and machine learning analysis of the MCR method for the Shopping domain

Target domain	Coefficients	stderr	R^2	R^2_{adj}	F	p	MSE
Shopping_d_addr	All	0.44	0.21	0.19	11.11	< 0.001	–
Shopping_d_addr	Location-all	0.46	0.10	0.10	14.54	< 0.001	–
Shopping_d_addr	Social-all	0.47	0.08	0.08	11.69	< 0.001	–
Shopping_d_addr	Mobile-all	0.45	0.16	0.16	24.62	< 0.001	–
Shopping_d_birth	All	0.39	0.22	0.20	11.70	< 0.001	–
Shopping_d_birth	Location-all	0.41	0.11	0.11	16.38	< 0.001	–
Shopping_d_birth	Social-all	0.41	0.10	0.09	13.35	< 0.001	–
Shopping_d_birth	Mobile-all	0.40	0.16	0.15	24.24	< 0.001	–
Shopping_d_name	All	0.40	0.10	0.07	4.42	0.01	–
Shopping_d_name	Location-all	0.41	0.04	0.03	5.39	0.02	–
Shopping_d_name	Social-all	0.41	0.06	0.06	8.81	0.00	–
Shopping_d_name	Mobile-all	0.41	0.05	0.05	7.31	0.01	–
Shopping_d_income	All	0.47	0.02	−0.00	0.90	0.44	–
Shopping_d_income	Location–all	0.47	0.01	0.00	1.40	0.24	–
Shopping_d_income	Social-all	0.47	0.01	0.00	1.26	0.26	–
Shopping_d_income	Mobile-all	0.47	0.02	0.01	1.96	0.16	–
Shopping_d_nutrition	All	0.44	0.15	0.13	7.55	< 0.001	–
Shopping_d_nutrition	Location-all	0.47	0.04	0.03	5.48	0.02	–
Shopping_d_nutrition	Social-all	0.45	0.08	0.08	11.57	< 0.001	–
Shopping_d_nutrition	Mobile-all	0.45	0.11	0.10	15.77	< 0.001	–
Shopping_d_allergy	All	0.41	0.13	0.11	6.11	< 0.001	–
Shopping_d_allergy	Location-all	0.42	0.03	0.02	4.21	0.04	–
Shopping_d_allergy	Social-all	0.42	0.07	0.06	9.08	0.00	–
Shopping_d_allergy	Mobile-all	0.41	0.09	0.09	13.24	< 0.001	–
Shopping_d_visit	All	0.45	0.14	0.12	6.95	< 0.001	–
Shopping_d_visit	Location-all	0.46	0.10	0.09	13.64	< 0.001	–
Shopping_d_visit	Social-all	0.47	0.05	0.05	7.32	0.01	–
Shopping_d_visit	Mobile-all	0.46	0.09	0.09	12.96	< 0.001	–
Shopping_d_whish	All	0.45	0.16	0.14	7.88	< 0.001	–
Shopping_d_whish	Location-all	0.46	0.09	0.08	11.84	< 0.001	–
Shopping_d_whish	Social-all	0.47	0.06	0.05	8.04	0.01	–
Shopping_d_whish	Mobile-all	0.45	0.12	0.11	17.42	< 0.001	–
Shopping_d_view	All	0.47	0.15	0.13	7.47	< 0.001	–
Shopping_d_view	Location-all	0.49	0.06	0.05	8.03	0.01	–
Shopping_d_view	Social-all	0.48	0.11	0.10	14.99	< 0.001	–
Shopping_d_view	Mobile-all	0.48	0.08	0.08	11.79	< 0.001	–
Shopping_d_receipt	All	0.50	0.03	0.01	1.29	0.28	–
Shopping_d_receipt	Location-all	0.50	0.03	0.02	3.40	0.07	–
Shopping_d_receipt	Social-all	0.50	0.02	0.01	2.27	0.13	–
Shopping_d_receipt	Mobile-all	0.50	0.00	−0.00	0.56	0.46	–
Shopping_d_cat	All	0.46	0.13	0.11	6.47	< 0.001	–

Table 24 (continued)

Target domain	Coefficients	stderr	R^2	R^2_{adj}	F	p	MSE
Shopping_d_cat	Location-All	0.47	0.07	0.07	10.10	0.00	–
Shopping_d_cat	Social-All	0.47	0.09	0.08	12.17	< 0.001	–
Shopping_d_cat	Mobile-all	0.47	0.07	0.06	9.78	0.00	–
Shopping_d_amount	All	0.48	0.07	0.05	3.22	0.02	–
Shopping_d_amount	Location-all	0.48	0.03	0.03	4.46	0.04	–
Shopping_d_amount	Social-all	0.49	0.02	0.01	2.88	0.09	–
Shopping_d_amount	Mobile-all	0.48	0.06	0.05	8.06	0.01	–

Table 25 Detailed validation study results for the statistical and machine learning analysis of the MCR method for the shopping domain continued

Target domain	Coefficients	stderr	R^2	R^2_{adj}	F	p	MSE
Shopping_d_price	All	0.49	0.07	0.05	3.19	0.03	–
Shopping_d_price	Location-all	0.49	0.02	0.02	3.25	0.07	–
Shopping_d_price	Social-all	0.49	0.03	0.03	4.50	0.04	–
Shopping_d_price	Mobile-all	0.49	0.05	0.05	7.32	0.01	–
Shopping_d_loyal	All	0.44	0.13	0.11	6.21	< 0.001	–
Shopping_d_loyal	Location-all	0.44	0.10	0.10	14.66	< 0.001	–
Shopping_d_loyal	Social-all	0.46	0.05	0.04	6.16	0.01	–
Shopping_d_loyal	Mobile-all	0.45	0.07	0.06	9.49	0.00	–
Shopping_d_loc	All	0.45	0.15	0.13	7.48	< 0.001	–
Shopping_d_loc	Location-all	0.46	0.11	0.10	15.86	< 0.001	–
Shopping_d_loc	Social-all	0.47	0.05	0.04	6.65	0.01	–
Shopping_d_loc	Mobile-all	0.46	0.09	0.09	13.30	< 0.001	–
Shopping_s_family	All	0.42	0.19	0.17	9.48	< 0.001	–
Shopping_s_family	Location-all	0.43	0.12	0.11	16.78	< 0.001	–
Shopping_s_family	Social-all	0.44	0.07	0.06	8.88	0.00	–
Shopping_s_family	Mobile-all	0.43	0.13	0.12	19.02	< 0.001	–
Shopping_s_friends	All	0.43	0.18	0.16	9.13	< 0.001	–
Shopping_s_friends	Location–all	0.45	0.11	0.10	15.45	< 0.001	–
Shopping_s_friends	Social–all	0.46	0.07	0.06	8.83	0.00	–
Shopping_s_friends	Mobile–all	0.44	0.13	0.12	18.86	< 0.001	–
Shopping_s_retailer	All	0.46	0.14	0.12	6.57	< 0.001	–
Shopping_s_retailer	Location-all	0.48	0.03	0.02	3.68	0.06	–
Shopping_s_retailer	Social-all	0.48	0.04	0.04	5.87	0.02	–
Shopping_s_retailer	Mobile-all	0.46	0.12	0.11	17.53	< 0.001	–
Shopping_s_third	All	0.47	0.04	0.02	1.88	0.14	–
Shopping_s_third	Location–all	0.47	0.03	0.02	3.91	0.05	–
Shopping_s_third	Social-all	0.47	0.00	−0.00	0.52	0.47	–
Shopping_s_third	Mobile-all	0.47	0.03	0.02	3.65	0.06	–

Table 26 Detailed validation study results for the statistical and machine learning analysis of the CCR method for the location domain

Target domain	Coefficients	stderr	R^2	R^2_{adj}	F	p	MSE
Location_t_famil	All	1.30	0.31	0.27	6.82	< 0.001	1.96
Location_t_famil	Social-all	1.40	0.16	0.14	11.61	< 0.001	2.14
Location_t_famil	Mobile-all	1.44	0.12	0.10	8.21	< 0.001	2.19
Location_t_famil	shopping-all	1.40	0.16	0.15	12.25	< 0.001	2.09
Location_t_event	All	1.18	0.37	0.31	6.20	< 0.001	1.72
Location_t_event	Social-all	1.25	0.23	0.22	19.24	< 0.001	1.75
Location_t_event	shopping-all	1.35	0.11	0.10	7.72	< 0.001	2.09
Location_t_movie	All	1.64	0.28	0.23	5.72	< 0.001	2.96
Location_t_movie	Social-all	1.69	0.20	0.18	10.38	< 0.001	3.60
Location_t_movie	shopping-all	1.80	0.11	0.07	3.05	0.012	3.73
Location_t_polit	All	1.63	0.21	0.15	3.89	< 0.001	3.19
Location_t_polit	Social-all	1.63	0.17	0.15	12.61	< 0.001	2.95
Location_t_polit	Mobile-all	1.75	0.03	0.02	2.19	0.116	3.25
Location_t_polit	shopping-all	1.71	0.09	0.07	4.14	0.008	2.84
Location_t_food	All	1.43	0.39	0.35	8.54	< 0.001	2.54
Location_t_food	Social-all	1.53	0.28	0.25	11.85	< 0.001	2.68
Location_t_food	shopping-all	1.63	0.16	0.15	12.29	< 0.001	2.77
Location_t_work	All	1.36	0.34	0.31	10.66	< 0.001	2.41
Location_t_work	Social-all	1.38	0.30	0.29	26.79	< 0.001	2.12
Location_t_hobbi	All	1.36	0.34	0.29	6.78	< 0.001	2.15
Location_t_hobbi	Social-all	1.45	0.21	0.20	16.92	< 0.001	2.34
Location_t_hobbi	Shopping-all	1.52	0.14	0.12	6.81	< 0.001	2.49
Location_t_music	All	1.62	0.37	0.32	7.67	< 0.001	2.94
Location_t_music	Social-all	1.68	0.28	0.27	16.48	< 0.001	2.95
Location_t_music	Mobile-all	1.87	0.10	0.09	7.38	< 0.001	3.88
Location_t_music	Shopping-all	1.89	0.10	0.08	4.58	0.004	3.97
Location_t_trave	All	1.30	0.29	0.25	6.95	< 0.001	2.18
Location_t_trave	Social-all	1.43	0.11	0.09	7.42	< 0.001	2.35
Location_t_sport	All	1.41	0.29	0.25	7.03	< 0.001	2.37
Location_t_sport	Social-all	1.45	0.22	0.20	17.45	< 0.001	2.44
Location_t_sport	Shopping-all	1.52	0.15	0.12	5.48	< 0.001	2.48
Location_t_tech	All	1.70	0.31	0.26	6.76	< 0.001	3.08
Location_t_tech	Shopping-all	1.85	0.15	0.13	7.31	< 0.001	3.95
Location_r_acqu	All	1.40	0.37	0.32	7.69	< 0.001	2.10
Location_r_acqu	Social-all	1.47	0.28	0.26	15.81	< 0.001	2.14
Location_r_school	All	1.42	0.26	0.22	5.38	< 0.001	2.09
Location_r_school	Social-all	1.46	0.20	0.18	7.85	< 0.001	2.14
Location_r_online	All	1.42	0.22	0.18	5.80	< 0.001	2.36
Location_r_online	Social-all	1.44	0.16	0.15	12.32	< 0.001	2.42
Location_r_online	Shopping-all	1.51	0.09	0.07	5.95	0.003	2.39
Location_r_sports	All	1.44	0.25	0.21	6.79	< 0.001	2.51

Table 26 (continued)

Target domain	Coefficients	stderr	R^2	R^2_{adj}	F	p	MSE
Location_r_sports	Mobile-all	1.55	0.09	0.08	6.60	0.002	2.82
Location_r_exfam	All	1.34	0.42	0.39	11.04	< 0.001	2.03
Location_r_exfam	Social-all	1.51	0.23	0.21	12.58	< 0.001	2.28
Location_r_imfam	All	1.43	0.35	0.29	6.34	< 0.001	2.38
Location_r_imfam	Social-all	1.52	0.22	0.20	8.89	< 0.001	2.55
Location_r_imfam	Mobile-all	1.60	0.13	0.11	9.27	< 0.001	2.73
Location_r_imfam	Shopping-all	1.52	0.23	0.20	7.47	< 0.001	2.80
Location_r_work	All	1.36	0.34	0.31	10.66	< 0.001	2.41
Location_r_work	Social-all	1.38	0.30	0.29	26.79	< 0.001	2.12
Location_r_closef	All	1.33	0.42	0.37	8.57	< 0.001	2.23
Location_r_closef	Social-all	1.41	0.33	0.29	9.91	< 0.001	2.37
Location_r_closef	Mobile-all	1.61	0.10	0.09	7.27	0.001	2.81
Location_r_closef	Shopping-all	1.53	0.20	0.18	10.21	< 0.001	2.83
Location_r_friends	All	1.33	0.40	0.37	11.74	< 0.001	1.93
Location_r_friends	Social-all	1.42	0.30	0.29	18.13	< 0.001	1.99
Location_r_friends	Mobile-all	1.54	0.17	0.16	13.20	< 0.001	2.34
Location_r_friends	Shopping-all	1.58	0.13	0.12	9.39	< 0.001	2.67

Table 27 Detailed validation study results for the statistical and machine learning analysis of the CCR method for the Social media domain

Target domain	Coefficients	stderr	R^2	R^2_{adj}	F	p	MSE
Social_t_famil	All	0.91	0.09	0.05	2.10	0.058	1.01
Social_t_famil	Mobile-all	0.93	0.03	0.02	2.23	0.112	0.97
Social_t_famil	Shopping-all	0.92	0.04	0.02	2.52	0.085	0.94
Social_t_event	All	0.72	0.27	0.22	6.26	< 0.001	0.59
Social_t_event	location-all	0.72	0.23	0.21	18.48	< 0.001	0.57
Social_t_event	Mobile-all	0.78	0.09	0.07	6.19	0.003	0.68
Social_t_event	Shopping-all	0.78	0.10	0.08	6.78	0.002	0.72
Social_t_movie	All	0.70	0.28	0.24	6.77	< 0.001	0.54
Social_t_movie	Location-all	0.68	0.30	0.29	27.54	< 0.001	0.49
Social_t_movie	Mobile-all	0.78	0.08	0.07	5.78	0.004	0.65
Social_t_movie	Shopping-all	0.78	0.11	0.08	3.68	0.007	0.66
Social_t_polit	All	0.93	0.38	0.33	8.09	< 0.001	1.07
Social_t_polit	Location-all	0.93	0.35	0.33	16.56	< 0.001	1.02
Social_t_food	Mobile-all	0.73	0.09	0.07	6.11	0.003	0.56
Social_t_work	All	0.71	0.37	0.33	10.17	< 0.001	0.55
Social_t_work	Location-all	0.72	0.31	0.30	28.76	< 0.001	0.54
Social_t_work	Shopping-all	0.81	0.14	0.13	10.26	< 0.001	0.69
Social_t_hobbi	All	0.66	0.20	0.17	6.15	< 0.001	0.50
Social_t_hobbi	Location-all	0.66	0.17	0.15	8.25	< 0.001	0.52
Social_t_hobbi	Mobile-all	0.70	0.07	0.06	4.79	0.010	0.54
Social_t_hobbi	Shopping-all	0.68	0.12	0.11	8.97	< 0.001	0.50
Social_t_music	All	0.69	0.18	0.15	5.38	< 0.001	0.60
Social_t_music	Location-all	0.68	0.20	0.17	6.17	< 0.001	0.55
Social_t_music	Mobile-all	0.74	0.03	0.01	1.22	0.305	0.63
Social_t_music	Shopping-all	0.72	0.09	0.05	2.30	0.049	0.65
Social_t_trave	All	0.74	0.25	0.19	4.36	< 0.001	0.70
Social_t_trave	Location-all	0.76	0.18	0.17	14.01	< 0.001	0.70
Social_t_trave	Mobile-all	0.81	0.06	0.04	2.63	0.053	0.80
Social_t_trave	Shopping-all	0.82	0.03	0.01	1.50	0.217	0.86
Social_t_sport	All	0.67	0.23	0.20	7.47	< 0.001	0.51
Social_t_sport	Location-all	0.67	0.22	0.20	11.60	< 0.001	0.52
Social_t_sport	Mobile-all	0.73	0.06	0.05	4.32	0.015	0.57
Social_t_sport	Shopping-all	0.71	0.11	0.10	8.15	< 0.001	0.53
Social_t_tech	All	0.85	0.21	0.15	3.44	< 0.001	0.79
Social_t_tech	Location-all	0.85	0.17	0.16	12.76	< 0.001	0.80
Social_t_tech	Mobile-all	0.91	0.04	0.03	2.92	0.057	0.91
Social_r_acqu	All	0.61	0.31	0.27	7.78	< 0.001	0.44
Social_r_acqu	Location-all	0.62	0.26	0.25	22.42	< 0.001	0.45
Social_r_acqu	Shopping-all	0.67	0.14	0.11	4.08	0.002	0.54
Social_r_school	All	0.58	0.22	0.19	6.91	< 0.001	0.41
Social_r_school	Location-all	0.60	0.16	0.14	7.88	< 0.001	0.42

Table 27 (continued)

Target domain	Coefficients	stderr	R^2	R^2_{adj}	F	p	MSE
Social_r_school	Shopping-all	0.61	0.12	0.10	5.77	< 0.001	0.45
Social_r_online	All	0.68	0.14	0.11	4.03	0.002	0.54
Social_r_online	Location-all	0.63	0.24	0.22	19.50	< 0.001	0.44
Social_r_online	Mobile-all	0.70	0.05	0.04	3.63	0.029	0.54
Social_r_online	Shopping-all	0.69	0.09	0.07	5.88	0.004	0.53
Social_r_sports	Location-all	0.79	0.18	0.16	9.22	< 0.001	0.71
Social_r_sports	Mobile-all	0.85	0.04	0.02	2.55	0.082	0.78
Social_r_sports	Shopping-all	0.82	0.11	0.10	7.93	< 0.001	0.73
Social_r_exfam	All	0.62	0.19	0.15	4.78	< 0.001	0.48
Social_r_exfam	Location-all	0.59	0.27	0.24	8.93	< 0.001	0.41
Social_r_exfam	Mobile-all	0.67	0.04	0.02	1.87	0.139	0.44
Social_r_exfam	Shopping-all	0.64	0.11	0.10	8.18	< 0.001	0.48
Social_r_imfam	All	0.58	0.22	0.18	5.00	< 0.001	0.42
Social_r_imfam	Shopping-all	0.62	0.11	0.08	3.75	0.006	0.40
Social_r_work	All	0.71	0.37	0.33	10.17	< 0.001	0.55
Social_r_work	Location-all	0.72	0.31	0.30	28.76	< 0.001	0.54
Social_r_work	Shopping-all	0.81	0.14	0.13	10.26	< 0.001	0.69
Social_r_closef	All	0.66	0.19	0.13	3.11	0.002	0.55
Social_r_closef	Location-all	0.63	0.23	0.21	9.50	< 0.001	0.46
Social_r_closef	Mobile-all	0.69	0.07	0.05	4.57	0.012	0.55
Social_r_friends	Mobile-all	0.73	0.04	0.03	2.83	0.063	0.59

Table 28 Detailed validation study results for the statistical and machine learning analysis of the CCR method for the mobile phone domain

Target domain	Coefficients	stderr	R^2	R^2_{adj}	F	p	AUC	F1
Mobile_c_comm	All	0.43	0.22	0.13	2.52	0.004	0.61	0.16
Mobile_c_comm	Location-all	0.47	0.01	−0.00	0.76	0.471	0.59	0.00
Mobile_c_comm	Social-all	0.45	0.07	0.04	2.94	0.036	0.65	0.08
Mobile_c_edu	All	0.44	0.14	0.05	1.59	0.103	0.57	0.04
Mobile_c_edu	Location-all	0.43	0.10	0.07	3.41	0.011	0.56	0.03
Mobile_c_edu	Social-all	0.44	0.07	0.05	3.15	0.027	0.66	0.21
Mobile_c_edu	Shopping-all	0.44	0.08	0.06	3.83	0.012	0.52	0.00
Mobile_c_news	All	0.45	0.21	0.18	6.60	< 0.001	0.74	0.52
Mobile_c_news	Location-all	0.46	0.17	0.15	8.53	< 0.001	0.68	0.51
Mobile_c_news	Social-all	0.50	0.02	0.00	1.03	0.382	0.56	0.17
Mobile_c_news	Shopping-all	0.47	0.12	0.10	5.57	0.001	0.61	0.42
Mobile_c_shopp	All	0.48	0.11	0.07	3.07	0.012	0.76	0.66
Mobile_c_shopp	Social-all	0.49	0.05	0.04	3.38	0.037	0.57	0.32
Mobile_c_shopp	Shopping-all	0.48	0.09	0.07	3.24	0.014	0.67	0.46
Mobile_c_Social	All	0.45	0.17	0.14	5.20	< 0.001	0.71	0.36
Mobile_c_Social	Location-all	0.45	0.15	0.14	11.07	< 0.001	0.67	0.36
Mobile_c_Social	Shopping-all	0.47	0.10	0.06	2.73	0.022	0.58	0.29
Mobile_c_tools	All	0.45	0.20	0.08	1.66	0.061	0.60	0.31
Mobile_c_tools	Location-all	0.46	0.09	0.03	1.65	0.127	0.49	0.03
Mobile_c_tools	Social-all	0.46	0.03	0.00	1.21	0.310	0.58	0.00
Mobile_c_travel	All	0.44	0.10	0.04	1.83	0.087	0.60	0.08
Mobile_c_travel	Location-all	0.43	0.09	0.07	5.92	0.003	0.47	0.27
Mobile_c_travel	Social-all	0.44	0.05	0.03	2.09	0.104	0.59	0.00
Mobile_c_travel	Shopping-all	0.44	0.04	0.01	1.26	0.291	0.55	0.11
Mobile_c_games	All	0.50	0.07	0.00	1.01	0.436	0.53	0.55
Mobile_c_games	Location-all	0.50	0.03	0.02	2.06	0.132	0.63	0.64
Mobile_c_games	Social-all	0.51	0.00	−0.02	0.01	0.999	0.34	0.65
Mobile_c_games	Shopping-all	0.50	0.04	−0.00	0.94	0.455	0.57	0.48

Table 29 Detailed validation study results for the statistical and machine learning analysis of the CCR method for the shopping domain

Target domain	Coefficients	stderr	R^2	R^2_{adj}	F	p	AUC	F1
Shopping_d_addr	All	0.45	0.22	0.16	4.11	< 0.001	0.66	0.56
Shopping_d_addr	Location-all	0.47	0.10	0.09	7.25	0.001	0.61	0.46
Shopping_d_addr	Social-all	0.47	0.08	0.07	5.83	0.004	0.54	0.27
Shopping_d_addr	Mobile-all	0.45	0.17	0.15	8.28	< 0.001	0.65	0.43
Shopping_d_birth	All	0.38	0.24	0.22	9.82	< 0.001	0.80	0.42
Shopping_d_birth	Location-all	0.41	0.12	0.11	8.54	< 0.001	0.62	0.07
Shopping_d_birth	Mobile-all	0.39	0.19	0.18	14.84	< 0.001	0.76	0.37
Shopping_d_name	All	0.41	0.11	0.05	2.03	0.057	0.52	0.19
Shopping_d_name	Location-all	0.41	0.05	0.03	3.03	0.052	0.61	0.00
Shopping_d_name	Social-all	0.41	0.07	0.06	4.76	0.010	0.68	0.24
Shopping_d_name	Mobile-all	0.41	0.07	0.05	4.66	0.011	0.59	0.04
Shopping_d_income	All	0.48	0.03	−0.02	0.54	0.775	0.48	0.79
Shopping_d_income	Mobile-all	0.46	0.07	0.04	2.31	0.061	0.61	0.81
Shopping_d_nutrition	All	0.45	0.16	0.12	3.38	0.002	0.65	0.29
Shopping_d_nutrition	Location-all	0.47	0.07	0.03	1.88	0.103	0.52	0.09
Shopping_d_nutrition	Social-all	0.46	0.08	0.07	5.85	0.004	0.58	0.07
Shopping_d_nutrition	Mobile-all	0.45	0.12	0.11	8.76	< 0.001	0.65	0.34
Shopping_d_allergy	All	0.41	0.13	0.10	3.72	0.004	0.65	0.00
Shopping_d_allergy	Location-all	0.42	0.06	0.04	2.89	0.038	0.64	0.00
Shopping_d_allergy	Mobile-all	0.41	0.09	0.08	6.58	0.002	0.54	0.08
Shopping_d_visit	All	0.45	0.15	0.11	4.25	0.001	0.71	0.47
Shopping_d_visit	Social-all	0.47	0.05	0.04	3.64	0.029	0.50	0.03
Shopping_d_whish	All	0.45	0.21	0.14	2.83	0.003	0.65	0.48
Shopping_d_whish	Location-all	0.47	0.09	0.06	3.21	0.015	0.62	0.37
Shopping_d_whish	Social-all	0.47	0.06	0.04	3.99	0.021	0.51	0.03
Shopping_d_whish	Mobile-all	0.45	0.14	0.13	10.34	< 0.001	0.64	0.23
Shopping_d_view	All	0.47	0.17	0.14	5.07	< 0.001	0.67	0.60
Shopping_d_view	Social-all	0.48	0.11	0.08	3.73	0.007	0.65	0.50
Shopping_d_view	Mobile-all	0.48	0.10	0.08	6.90	0.001	0.63	0.59
Shopping_d_receipt	All	0.50	0.08	−0.01	0.89	0.550	0.60	0.66
Shopping_d_receipt	Location-all	0.49	0.05	0.03	3.28	0.041	0.58	0.63
Shopping_d_receipt	Mobile-all	0.50	0.01	−0.02	0.33	0.802	0.54	0.57
Shopping_d_cat	All	0.46	0.18	0.12	3.19	0.003	0.56	0.32
Shopping_d_cat	Location-all	0.48	0.09	0.05	2.36	0.044	0.57	0.36
Shopping_d_cat	Social-all	0.47	0.10	0.07	4.43	0.005	0.59	0.31
Shopping_d_amount	All	0.47	0.15	0.10	2.73	0.008	0.66	0.32
Shopping_d_amount	Location-all	0.48	0.08	0.05	2.61	0.039	0.59	0.16
Shopping_d_amount	Social-all	0.49	0.03	0.02	2.05	0.134	0.56	0.03
Shopping_d_amount	Mobile-all	0.48	0.07	0.05	4.40	0.014	0.63	0.41
Shopping_d_price	All	0.48	0.13	0.07	2.21	0.031	0.60	0.44
Shopping_d_price	Location-all	0.49	0.03	0.01	1.91	0.153	0.50	0.17

Table 29 (continued)

Target domain	Coefficients	stderr	R^2	R^2_{adj}	F	p	AUC	F1
Shopping_d_price	Mobile-all	0.49	0.06	0.04	3.95	0.022	0.57	0.48
Shopping_d_loyal	All	0.45	0.16	0.09	2.32	0.016	0.72	0.42
Shopping_d_loyal	Location-all	0.44	0.12	0.10	8.27	< 0.001	0.66	0.27
Shopping_d_loyal	Social-all	0.46	0.06	0.03	2.50	0.063	0.46	0.00
Shopping_d_loyal	Mobile-all	0.45	0.10	0.08	4.54	0.005	0.64	0.09
Shopping_d_loc	All	0.46	0.17	0.11	3.06	0.004	0.60	0.43
Shopping_d_loc	Location-all	0.46	0.11	0.08	3.97	0.005	0.66	0.37
Shopping_d_loc	Social-all	0.48	0.06	0.03	2.46	0.066	0.47	0.03
Shopping_d_loc	Mobile-all	0.46	0.11	0.09	5.25	0.002	0.66	0.40
Shopping_s_family	All	0.42	0.19	0.17	7.34	< 0.001	0.64	0.45
Shopping_s_family	Location-all	0.43	0.13	0.12	9.44	< 0.001	0.60	0.27
Shopping_s_family	Social-all	0.45	0.07	0.04	2.46	0.049	0.69	0.16
Shopping_s_family	Mobile-all	0.43	0.13	0.12	9.46	< 0.001	0.61	0.03
Shopping_s_retailer	All	0.45	0.21	0.14	2.90	0.002	0.61	0.23
Shopping_s_retailer	Location-all	0.48	0.04	0.00	1.11	0.359	0.52	0.03
Shopping_s_retailer	Social-all	0.48	0.06	0.02	1.48	0.201	0.52	0.14
Shopping_s_retailer	Mobile-all	0.45	0.14	0.12	10.02	< 0.001	0.67	0.48
Shopping_s_third	All	0.47	0.07	0.01	1.25	0.281	0.44	0.77
Shopping_s_third	Location-all	0.47	0.03	−0.00	0.96	0.433	0.51	0.80
Shopping_s_third	Mobile-all	0.47	0.05	0.01	1.26	0.285	0.54	0.77

Funding Open Access funding enabled and organized by Projekt DEAL.

References

Addae, Joyce H. et al.: "Exploring user behavioral data for adaptive cybersecurity". In: User Modeling and User-Adapted Interaction 29.3 (2019), pp. 701–750. https://doi.org/10.1007/s11257-019-09236-5

Adomavicius, G., Tuzhilin, A.: "Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions". In: IEEE Transactions on Knowledge and Data Engineering 17.6 (2005), pp. 734–749. https://doi.org/10.1109/TKDE.2005.99

Barnes, Susan B.: "A privacy paradox: Social networking in the United States." In: First Monday 11.9 (2006)

Benisch, M. et al.:"Capturing Location-privacy Preferences: Quantifying Accuracy and User-burden Tradeoffs". In: Personal Ubiquitous Comput. 15.7 (2011), pp. 679–694. issn: 1617-4909. https://doi.org/10.1007/s00779-010-0346-0

Buhrmester, Michael, Kwang, Tracy, Gosling, Samuel D.: 'Amazon's Mechanical Turk: A new Source of Inexpensive, Yet High-Quality, Data?" In: Perspectives on Psychological Science 6.1 (2011), pp. 3–5. issn: 1745- 6916. https://doi.org/10.1177/1745691610393980

Christin, Delphine, Michalak, Martin, Hollick, Matthias: "Raising User Awareness About Privacy Threats in Participatory Sensing Applications Through Graphical Warnings". In: Proceedings of International Conference on Advances in Mobile Computing #38; Multimedia. MoMM '13. Vienna, Austria: ACM, (2013), 445:445–445:454. isbn: 978-1-4503- 2106-8. https://doi.org/10.1145/2536853.2536861

Connelly, Kay, Khalil, Ashraf, Liu, Yong: "Do I Do What I Say?: Observed Versus Stated Privacy Preferences". In: Proceedings of the 11th IFIP TC 13 International Conference on Human-computer Interaction. INTERACT'07. Rio de Janeiro, Brazil: Springer-Verlag, (2007), pp. 620– 623. isbn: 3-540-74794-X, 978-3-540-74794-9. http://dl.acm.org/citation.cfm?id=1776994.1777074

Consolvo, Sunny, et al.: "Location Disclosure to Social Relations: Why, when, & What People Want to Share". In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI '05. Portland, Oregon, USA: ACM, (2005), pp. 81–90. isbn: 1-58113-998-5. https://doi.org/10.1145/1054972.1054985

Costa, P.T., McCrae, R.R.: Inc Psychological Assessment Resources. Revised NEO Personality Inventory (NEO PI-R) and NEO Five-Factor Inventory (NEO-FFI). Psychological Assessment Resources, (1992). https://books.google.de/books?id=mp3zNwAACAAJ

Fabian, A., et al.: "Cross-system user modeling and personalization on the Social Web". In: User Modeling and User-Adapted Interaction 23.2 (2013), pp. 169–209. issn: 1573-1391. https://doi.org/10.1007/s11257-012-9131-2

Das, Sauvik, Dabbish, Laura A., Hong, Jason I.: "A Typology of Perceived Triggers for End-User Security and Privacy Behaviors". In: Proceedings of the Fifteenth USENIX Conference on Usable Privacy and Security. SOUPS'19. Santa Clara, CA, USA: USENIX Association, (2019), pp. 97–115. isbn: 9781939133052

Dey, R., Jelveh, Z., Ross, K.: "Facebook users have become much more private: A large-scale study". In: 2012 IEEE International Conference on Pervasive Computing and Communications Workshops. (2012), pp. 346– 352. https://doi.org/10.1109/PerComW.2012.6197508

Ebada Mohamed, Reham, Chiasson, Sonia: "Online Privacy and Aging of Digital Artifacts". In: Proceedings of the Fourteenth USENIX Confer- ence on Usable Privacy and Security. SOUPS '18. Baltimore, MD, USA: USENIX Association, (2018), pp. 177–195. isbn: 9781931971454

Ebert, Nico, Alexander Ackermann, Kurt, Heinrich, Peter: 'Does Context in Privacy Communication Really Matter? | A Survey on Consumer Concerns and Preferences". In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. CHI '20. Honolulu, HI, USA: Association for Computing Machinery, (2020), pp. 1–11. isbn: 9781450367080. https://doi.org/10.1145/3313831.3376575

Ebert, Nico, Alexander Ackermann, Kurt, Scheppler, Björn: "Bolder is Better: Raising User Awareness through Salient and Concise Privacy Notices". In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. CHI '21. Yokohama, Japan: Association for Computing Machinery, (2021). isbn: 9781450380966. https://doi.org/10.1145/3411764.3445516

Faklaris, Cori, Dabbish, Laura, Hong, Jason I.: "A Self-Report Measure of End-User Security Attitudes (SA-6)". In: Proceedings of the Fif- teenth USENIX Conference on Usable Privacy and Security. SOUPS'19. Santa Clara, CA, USA: USENIX Association, (2019), pp. 61–77. isbn: 9781939133052

Fang, Lujun, LeFevre, Kristen: "Privacy Wizards for Social Networking Sites". In: Proceedings of the 19th International Conference on World Wide Web. WWW '10. Raleigh, North Carolina, USA: ACM, (2010), pp. 351–360. isbn: 978-1-60558-799-8. https://doi.org/10.1145/1772690.1772727

Farnadi, Golnoosh, et al.: "Computational personality recognition in social media". In: User Modeling and User-Adapted Interaction 26.2 (2016), pp. 109–142. issn: 1573-1391. https://doi.org/10.1007/s11257-016-9171-0

Fisher, Ronald A.: The Design of Experiments (9th ed.) Macmillan, (1971)

Friedman, Arik, Berkovsky, Shlomo, Ali Kaafar, Mohamed: "A differential privacy framework for matrix factorization recommender systems". In: User Modeling and User-Adapted Interaction 26.5 (2016), pp. 425–458. issn: 1573-1391. https://doi.org/10.1007/s11257-016-9177-7

Ghaiumy Anaraky, Reza, et al.: "To Disclose or Not to Disclose: Examining the Privacy Decision-Making Processes of Older vs. Younger Adults". In: Proceedings of the 2021 CHI Conference on Human Factors in Com- puting Systems. CHI '21. Yokohama, Japan: Association for Computing Machinery, (2021). isbn: 9781450380966. https://doi.org/10.1145/3411764.3445204

Gross, Ralph, Acquisti, Alessandro: "Information Revelation and Privacy in Online Social Networks". In: Proceedings of the 2005 ACM Work- shop on Privacy in the Electronic Society. WPES '05. Alexandria, VA, USA: ACM, (2005), pp. 71–80. isbn: 1-59593-228-3. https://doi.org/10.1145/1102199.1102214

Heckmann, D.: Ubiquitous User Modeling. Dissertationen zur künstlichen Intelligenz - DISKI. Akademische Verlagsgesellschaft, (2006). isbn: 9783898382977. https://books.google.de/books?id=e5adLEi4gYgC

Hutton, Luke, Henderson, Tristan, Kapadia, Apu: "Short paper: "here i am, now pay me!": privacy concerns in incentivised location-sharing systems". In: 7th ACM Conference on Security & Privacy in Wireless and Mobile Networks, WiSec'14, Oxford, United Kingdom, July 23-25, 2014. (2014), pp. 81–86. https://doi.org/10.1145/2627393.2627416

Ismail, Qatrunnada, et al.: "Crowdsourced Exploration of Security Con- figurations". In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. CHI '15. Seoul, Republic of Korea: ACM, (2015), pp. 467–476. isbn: 978-1-4503-3145-6. https://doi.org/10.1145/2702123.2702370

Knijnenburg, Bart P., Kobsa, Alfred, Jin, Hongxia: "Dimensionality of Information Disclosure Behavior". In: Int. J. Hum.-Comput. Stud. 71.12 (2013), pp. 1144–1162. issn: 1071-5819. https://doi.org/10.1016/j.ijhcs.2013.06.003

Liu, Bin, Lin, Jialiu, Sadeh, Norman: "Reconciling Mobile App Privacy and Usability on Smartphones: Could User Privacy Profiles Help?" In: Proceedings of the 23rd International Conference on World Wide Web. WWW '14. Seoul, Korea: ACM, (2014), pp. 201–212. isbn: 978-1- 4503-2744-2. https://doi.org/10.1145/2566486.2568035

Liu, Bin, et al.: "Follow My Recommendations: A Personalized Privacy Assistant for Mobile App Permissions". In: Twelfth Symposium on Us- able Privacy and Security (SOUPS 2016). Denver, CO: USENIX Association, June (2016), pp. 27–41. isbn: 978-1-931971-31-7. https: //www.usenix.org/conference/soups2016/technical- sessions/presentation/liu

Lugano, Giuseppe, Saariluoma, Pertti: "To Share or Not to Share: Supporting the User Decision in Mobile Social Software Applications". English. In: User Modeling 2007. Ed. by Cristina Conati, Kathleen Mc- Coy, and Georgios Paliouras. Vol. 4511. Lecture Notes in Computer Science.

Springer Berlin Heidelberg, (2007), pp. 440–444. isbn: 978-3-540- 73077-4. https://doi.org/10.1007/978-3-540-73078-1_61

Lynn Dupree, Janna, et al.: "Privacy Personas: Clustering Users via Attitudes and Behaviors toward Security Practices". In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. CHI '16. San Jose, California, USA: Association for Computing Machinery, (2016), pp. 5228–5239. isbn: 9781450333627. https://doi.org/10.1145/2858036.2858214

Majeski, Michelle, Johnson, Maritza, Bellovin, Steven M.: The Fail- ure of Online Social Network Privacy Settings. Tech. rep. CUCS-010-11. Department of Computer Science, Columbia University, Feb. (2011)

Malhotra, Naresh K., Kim, Sung S., Agarwal, James: "Internet Users' Information Privacy Concerns (IUIPC): The Construct, the Scale, and a Causal Model". In: Info. Sys. Research 15.4 (2004), pp. 336– 355. issn: 1526-5536. https://doi.org/10.1287/isre.1040.0032

Mondal, Mainack, et al.: "Understanding and Specifying Social Access Control Lists". In: Symposium On Usable Privacy and Security (SOUPS 2014). Menlo Park, CA: USENIX Association, July (2014), pp. 271–283. isbn: 978-1-931971-13-3. https://www.usenix.org/conference/ soups2014/proceedings/presentation/mondal

Murillo, Ambar, et al.: ""If I Press Delete, It's Gone": User Understanding of Online Data Deletion and Expiration". In: Proceedings of the Four- teenth USENIX Conference on Usable Privacy and Security. SOUPS '18. Baltimore, MD, USA: USENIX Association, (2018), pp. 329–339. isbn: 9781931971454

Young Park, Cheul, et al.: "Share and Share Alike? An Exploration of Secure Behaviors in Romantic Relationships". In: Proceedings of the Four- teenth USENIX Conference on Usable Privacy and Security. SOUPS '18. Baltimore, MD, USA: USENIX Association, (2018), pp. 83–102. isbn: 9781931971454

Patil, Sameer et al.: "My Privacy Policy: Exploring End-user Specification of Free-form Location Access Rules". In: Financial Cryptography and Data Security: FC 2012 Workshops, USEC and WECSR 2012, Kral- endijk, Bonaire, March 2, 2012, Revised Selected Papers. Ed. by Jim Blyth, Sven Dietrich, and L. Jean Camp. Berlin, Heidelberg: Springer Berlin Heidelberg, (2012), pp. 86–97. isbn: 978-3-642-34638-5. https://doi.org/10.1007/978-3-642-34638-5_8

Paul, Thomas, Puscher, Daniel, Strufe, Thorsten: "Improving the Usability of Privacy Settings in Facebook". In: CoRR abs/1109.6046 (2011). arXiv:1109.6046

Raber, F., Krüger, A.: "Deriving Privacy Settings for Location Sharing: Are Context Factors Always the Best Choice?" In, : IEEE Sym- posium on Privacy-Aware Computing (PAC). Sept. **2018**, 86–94 (2018). https://doi.org/10.1109/PAC.2018.00015

Raber, Frederic, Krüger, Antonio: "Privacy Perceiver: Using Social Network Posts to Derive Users' Privacy Measures". In: Adjunct Publi- cation of the 26th Conference on User Modeling, Adaptation and Per- sonalization. UMAP '18. Singapore, Singapore: ACM, (2018), pp. 227– 232. isbn: 978-1-4503-5784-5. https://doi.org/10.1145/3213586.3225228

Raber, Frederic, Krüger, Antonio: "Towards Understanding the Influence of Personality on Mobile App Permission Settings". In: Human- Computer Interaction - INTERACT 2017. IFIP Conference on Human- Computer Interaction (INTERACT-17), 16th IFIP TC 13 International Conference, September 25-29, Mumbai, India. Ed. by R. Bernhaupt et al. IFIP. Springer, (2017). isbn: 978-3-319-22698-9

Raber, Frederic, Ziemann, David, Krüger, Antonio: "The 'Retailio' Privacy Wizard: Assisting Users with Privacy Settings for Intelligent Retail Stores". In: EuroUSEC '18 : 3rd European Workshop on Usable Se- curity. EuroUSEC European Workshop on Usable Security (EuroUSEC- 18), 3rd, located at IEEE Conference on Security & Privacy, April 23, London, UCL, United Kingdom. Ed. by CharlesWeir and Michelle Mazurek. Internet Society, (2018). isbn: https://doi.org/10.14722/eurousec.2018.23009

Raber, Frederic, et al.: "Fine-grained Privacy Setting Prediction using a Privacy Attitude Questionnaire and Machine Learning". In: Human- Computer Interaction - INTERACT 2017. IFIP Conference on Human- Computer Interaction (INTERACT-17), 16th IFIP TC 13 International Conference, September 25-29, Mumbai, India. Ed. by R. Bernhaupt et al. IFIP. Springer, (2017). isbn: 978-3-319-22698-9

Rashidi, Yasmeen, et al.: ""You Don'tWant to Be the next Meme": College Students' Workarounds to Manage Privacy in the Era of Pervasive Photography". In: Proceedings of the Fourteenth USENIX Conference on Us- able Privacy and Security. SOUPS '18. Baltimore, MD, USA: USENIX Association, (2018), pp. 143–157. isbn: 9781931971454

Ravichandran, Ramprasad et al.: "Capturing Social Networking Privacy Preferences: Can Default Policies Help Alleviate Tradeoffs Between Expressiveness and User Burden?" In: Proceedings of the 5th Symposium on Usable Privacy and Security. SOUPS '09. Mountain View, California, USA: ACM, (2009), 47:1–47:1. isbn: 978-1-60558-736-3. https://doi.org/10.1145/1572532.1572587

Sahebi, Shaghayegh, Brusilovsky, Peter: "Cross-Domain Collaborative Recommendation in a Cold-Start Context: The Impact of User Profile Size on the Quality of Recommendation". In: User Modeling, Adaptation, and Personalization. Ed. by Sandra Carberry et al. Berlin, Heidelberg: Springer Berlin Heidelberg, (2013), pp. 289–295. isbn: 978-3-642-38844-6

Ref Sanchez, Odnan, et al.: "A recommendation approach for user privacy preferences in the fitness domain". In: User Modeling and User-Adapted Interaction 30 (2020). https://doi.org/10.1007/s11257-019-09246-3

Sannon, Shruti, Bazarova, Natalya N., Cosley, Dan: "Privacy Lies: Understanding How, When, and Why People Lie to Protect Their Privacy in Multiple Online Contexts". In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. CHI '18. Montreal QC, Canada: Association for Computing Machinery, (2018), pp. 1– 13. isbn: 9781450356206. https://doi.org/10.1145/3173574.3173626

Schein, Andrew I., et al.: "Methods and Metrics for Cold-start Recommendations". In: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '02. Tampere, Finland: ACM, (2002), pp. 253–260. isbn: 1-58113- 561-0. https://doi.org/10.1145/564376.564421

Shehab, Mohamed, Touati, Hakim: "Semi-Supervised Policy Recommendation for Online Social Networks". In: Proceedings of the 2012 In- ternational Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012). ASONAM '12. Washington, DC, USA: IEEE Computer Society, (2012), pp. 360–367. isbn: 978-0-7695-4799-2. https://doi.org/10.1109/ASONAM.2012.66

Sinha, Arunesh, Li, Yan, Bauer, Lujo: "What You Want is Not What You Get: Predicting Sharing Policies for Text-based Content on Facebook". In: Proceedings of the 2013 ACM Workshop on Artificial Intelli- gence and Security. AISec '13. Berlin, Germany: ACM, (2013), pp. 13–24. isbn: 978-1-4503-2488-5. https://doi.org/10.1145/2517312.2517317

Stutzman, Fred, Gross, Ralph, Acquisti, Alessandro: "Silent Listeners: The Evolution of Privacy and Disclosure on Facebook". In: Journal of Privacy and Confidentiality 4.2 (2013). https://doi.org/10.29012/jpc.v4i2.620. https://journalprivacyconfidentiality.org/index.php/jpc/article/view/620

Shyam Sundar, S., et al.: "Online Privacy Heuristics That Predict Information Disclosure". In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. CHI '20. Honolulu, HI, USA: Association for Computing Machinery, (2020), pp. 1–12. isbn: 9781450367080. https://doi.org/10.1145/3313831.3376854

Theil, Henri: "Best Linear Unbiased Estimation and Prediction". In: Principles of Econometrics (1971), pp. 119–124. issn: 0-471-85845-5

Toch, Eran, Wang, Yang, Faith Cranor, Lorrie: "Personalization and privacy: a survey of privacy risks and remedies in personalization-based systems". In: User Modeling and User-Adapted Interaction 22.1 (2012), pp. 203–220. issn: 1573-1391. https://doi.org/10.1007/s11257-011-9110-z

Tsai, Lynn, et al.: "Turtleguard: Helping Android Users Apply Contextual Privacy Preferences". In: Proceedings of the Thirteenth USENIX Con- ference on Usable Privacy and Security. SOUPS '17. Santa Clara, CA, USA: USENIX Association, (2017), pp. 145–162. isbn: 9781931971393

Wang, Yang, Kobsa, Alfred: "A PLA-based Privacy-enhancing User Modeling Framework and Its Evaluation". In: User Modeling and User- Adapted Interaction 23.1 (2013), pp. 41–82. issn: 0924-1868. https://doi.org/10.1007/s11257-011-9114-8

Watson, Jason, Richter Lipford, Heather, Besmer, Andrew: "Mapping User Preference to Privacy Default Settings". In: ACM Trans. Comput.-Hum. Interact. 22.6 (2015). issn: 1073-0516. https://doi.org/10.1145/2811257

Wijesekera, Primal, et al.: "Contextualizing Privacy Decisions for Better Prediction (and Protection)". In: Proceedings of the 2018 CHI Confer- ence on Human Factors in Computing Systems. CHI '18. Montreal QC, Canada: Association for Computing Machinery, (2018), pp. 1–13. isbn: 9781450356206. https://doi.org/10.1145/3173574.3173842

Wisniewski, Pamela J., Knijnenburg, Bart P., Richter Lipford, Heather: "Making Privacy Personal". In: Int. J. Hum.-Comput. Stud. 98.C (2017), pp. 95–108. issn: 1071-5819. https://doi.org/10.1016/j.ijhcs.2016.09.006

Wisniewski, Pamela, et al.: “Give Social Network Users the Privacy They Want". In: Proceedings of the 18th ACM Conference on Computer Sup- ported Cooperative Work & Social Computing. CSCW ’15. Vancouver, BC, Canada: ACM, (2015), pp. 1427–1441. isbn: 978-1-4503-2922-4. https://doi.org/10.1145/2675133.2675256

Wu, Wen, Chen, Li, Zhao, Yu: “Personalizing recommendation diversity based on user personality”. In: User Modeling and User-Adapted Interaction 28.3 (2018), pp. 237–276. issn: 1573-1391. https://doi.org/10.1007/s11257-018-9205-x

Zhou, Huiyuan, et al.: “Enhancing Mobile Content Privacy with Proxemics Aware Notifications and Protection”. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. CHI ’16. San Jose, California, USA: Association for Computing Machinery, (2016), pp. 1362–1373. isbn: 9781450333627. https://doi.org/10.1145/2858036.2858232

Zou, Yixin, et al.: “Examining the Adoption and Abandonment of Security, Privacy, and Identity Theft Protection Practices”. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. CHI ’20. Honolulu, HI, USA: Association for Computing Machinery, (2020), pp. 1–15. isbn: 9781450367080. https://doi.org/10.1145/3313831.3376570

Frederic Raber has a PhD in computer science in the field of “AI and data protection” from the Saarland University and is working at the German Research Center for Artificial Intelligence as a senior researcher. Through many working years of work at DFKI, he had contact with many areas of AI, from artificial intelligence in retail through intelligent systems for supporting workers in production to AI-based recommendation systems for data protection settings. Today, he uses his expertise to play a leading role in various projects for the introduction and implementation of AI and digitization topics.

Antonio Krüger is a CEO and scientific director of the German Research Center for Artificial Intelligence GmbH (DFKI) and head of the department “Cognitive Assistants” at DFKI. He is a full professor for Computer Science at Saarland University (since 2009), Head of the Ubiquitous Media Technology Lab and scientific director of the Innovative Retail Laboratory (IRL) at DFKI. From 2004 to 2009, he was a professor of computer science and geoinformatics at the University of Münster and acted as the managing director of the institute for geoinformatics. He studied computer science and economics at Saarland University and finished his PhD in 1999 as a member of the Saarbrücken graduate school of Cognitive Science. Antonio has published more than 200 scientific articles and papers in internationally recognized journals and conferences and is member of several steering committees, editorial boards, and scientific advisory committees.