Deriving privacy settings for location sharing: Are context factors always the best choice?

Frederic Raber
DFKI, Saarland Informatics Campus
frederic.raber@dfki.de

Antonio Krueger
DFKI, Saarland Informatics Campus
krueger@dfki.de

***Abstract*—Research has observed context factors like occasion and time as influential factors for predicting whether or not to share a location with online friends. In other domains like social networks, personality was also found to play an important role. Furthermore, users are seeking a fine-grained disclosement policy that also allows them to display an obfuscated location, like the center of the current city, to some of their friends. In this paper, we observe which context factors and personality measures can be used to predict the correct privacy level out of seven privacy levels, which include obfuscation levels like center of the street or current city. Our results show that a prediction is possible with a precision 20% better than a constant value. We will give design indications to determine which context factors should be recorded, and how much the precision can be increased if personality and privacy measures are recorded using either a questionnaire or automated text analysis.**

Keywords*—*privacy, machine learning, location sharing, user modeling, adaptation

I. Introduction

In modern society, location sharing is an integral part of a social network user's everyday activity: Whenever a post is created about a visited restaurant, event or a party with friends, users have the option to attach their current location to the post. While only 11% of the users took advantage of this functionality in 2013, this percentage had already doubled two years later. Using location sharing services on smartphones is even more common: More than 70% of smartphone users use their devices to share their location online or use it for a mapping service like Google Maps[1].

Research in the past has indicated that users tend to hide all information if the rule system provided by the location sharing provider is too simplistic and does not allow a fine-grained control of the location disclosure. Users then tend to hide all information rather than taking the risk of an unwanted disclosure[2]. Nevertheless, manually creating a set of rules is done by hardly any users, as it takes a lot of effort and technical knowledge of the consequences of each setting[32]. Researchers therefore began to automatically *infer* privacy rules by, for example, using user feedback, or location sharing decisions made in the past [32]. Unfortunately, if the user has never used the sharing functionality before, there is no data available for a prediction. Even worse, related work has shown that online privacy decisions do not correlate with the actual privacy desires of the user[1], making it necessary to explore different sources for the prediction. The current state of the art uses either context factors like the occasion or event where the location is shared[10] or the group of recipients[11], or individual factors like the personality or privacy concerns of the user [26] for predicting whether to disclose or not disclose the location. Related work has shown that these two disclosure levels are not sufficient: Most users wish to have a fine-grained disclosure policy that also allows them to share an obfuscated position like the street or city center instead of the exact location [25]. Using these obfuscation techniques, it is still possible, for example, to tell your friends that you are going on a date in a town nearby without disclosing the home address and the identity of the person you are dating. Also, social network providers like Facebook recently introduced the possibility to share obfuscated locations like only the name of the city, supporting these findings. Nevertheless, the users are left alone to decide on using obfuscated locations without any clarification about the usefulness and consequences for the user's privacy depending on the obfuscation level. Every obfuscation level comes with increased privacy, but decreased usefulness for the friends receiving the location. The user sharing the location therefore always has to make a tradeoff between privacy and usefulness.

None of the related work allows one to infer privacy rules that depend on both context factors and individual factors, including personality and privacy concerns, and that offer more than two location sharing levels, including obfuscated locations in addition to the exact or no GPS position. Our approach derives a location privacy setting that is tailored to the user, by combining personality and context factors, and allows a more precise definition of the privacy preference by offering seven distinct privacy levels. It does not require technical knowledge to create privacy rules, but uses a short and simple questionnaire as input to perform the derivation of privacy rules. Instead of filling out the questionnaire, it is also possible to use the posts written by the user to extract his personality and privacy measures out of his writing style, as related work has shown [7]. By this means, no additional user effort is needed to propose the privacy settings that are, based on our experiment results, clearly better than using a standard setting (later called a "constant value").

In detail, we try to solve the following research questions:

1) Is it possible to use personality/privacy concerns to predict fine-grained location sharing settings?
2) If fine-grained location privacy levels, including obfuscated locations in addition to a disclose/undisclose option are offered, are they used by the users? How often are obfuscated locations chosen instead of the

[1] http://www.pewinternet.org/2013/09/12/location-based-services/

exact or no location?

3) How good is the prediction performance when using the data of a dedicated privacy questionnaire, compared to using personality measures that can be automatically extracted?
4) How precisely can privacy rules be derived, if only context factors are taken into account?
5) Which variables other than personality and privacy concerns influence the privacy decision?

We conducted an online user study with 100 participants to shed light on these questions, and use a correlation and regression analysis to investigate *whether* there are correlations, and *how precisely* a prediction can be made using the recorded data. Furthermore, we explore which context factors influence the privacy decision, and how large the influence on the result is. Although predicting seven distinct privacy levels is surely harder than only deciding whether to disclose or not, we achieve a prediction that is about *20%* better than taking a constant value. The discussion section will finally give some design guidelines on which features should be used depending on the available data (context factors or personality or privacy measures) and the desired additional user interaction for filling out the questionnaires. In addition to a solution with a best possible precision, we also outline a solution that is optimized towards the least additional user effort.

II. Related work

The related work that is of interest for this paper can be divided into three sections: privacy and personality questionnaires that currently exist and that can be used to measure the user's personality and privacy concerns, other context factors that influence the privacy decision in location sharing, and lastly, other similar location privacy management tools that exist in research.

A. Privacy and personality questionnaires

Alan Westin's work [21] on consumer privacy indices is said to be the first publication that introduces a questionnaire to capture consumers' *privacy concerns*. Westin differentiates between three types of consumers: The *Unconcerned* hardly care about privacy and tend to give away all private data without any concerns. The opposite privacy type is called a *Fundamentalist*, who tries to disclose as little information as possible. The third group of consumers, the *Pragmatists*, tend to keep a balance between privacy and usability. They believe that privacy is important, but also accept the necessity to share information in order to benefit, for example, from bonus programs or tailored ads.

The Westin scales allow only a rough estimation of a general privacy attitude and are very coarse-grained. It is hard to use the privacy index of a user to predict behavior when it comes to privacy decisions, as the members of the privacy categories do not behave significantly differently[35]. In particular the categorization into only three privacy categories makes it hard to predict the user's reactions to hypothetical scenarios or permission settings. The authors point out that the questionnaire is too unspecific to capture any significant effects on privacy behavior.

The more detailed PCS[2] questionnaire [5] consists of 28 questions that result in three measures: General Caution, Technical Protection and Privacy Concern. Although more detailed than the 12-item Westin questionnaire, it still adresses the general level of privacy concern of a person; furthermore, it includes technical questions (for example about shredding floppy disks and CDs because of privacy issues) which seem outdated nowadays.

In contrast to the previous mentioned questionnaires, the CFIP[3] [33], and the IUIPC [23] questionnaire based on it, are tailored especially for privacy in the field of online shopping. The authors found that privacy concerns regarding online companies can be expressed using three measures: the *control* measure, which determines how far a subject desires to have control over the disclosure and transfer of her personal information, the desired *awareness* of how and to whom the personal information is disclosed, and *collection*, describing how important it is for the subject to know which personal data is collected. The IUIPC is, aside from of the Westin privacy indices, one of the most frequently used questionnaires and is also used in several related papers to predict privacy settings on Facebook[26], app permissions [27] or the disclosement settings for the data of an intelligent retail store [29], [30]. In our work, we will use both the traditional Westin questionnaire as well as the modern IUIPC questionnaire with its three measures control, awareness and collection for the later analyses. Recent work has shown that the IUIPC measures can also be derived from blog or social network entries [28] of a user. The user burden for gathering the big five personality measures can therefore be reduced to a minimum using machine-learning techniques.

The current research standard for capturing the *personality* of a user is the big five personal inventory (Costa and McCrae [12]). Although it was iteratively developed over several years and receives very positive reviews [19], [12], there are also some criticized aspects of the scale [3]. Nevertheless, it is established as the standard personal inventory questionnaire. The full version of the big five (also called the NEO-PI-R) in its newest form consists of 240 items, providing five personality measures: *Openness to experience*, denoting general appreciation for cultural aspects like art, emotions, adventures etc.; *Conscientiousness*, meaning the tendency to be self-disciplined; *Extraversion*, describing the level of social engagement; *Agreeableness*, meaning the ability to accommodate and cooperate with other people and *Neuroticism*, as the tendency to experience and express negative feelings. The questionnaire in its original version requires up to 30 to 40 minutes to fill out, making it unsuitable in most scenarios. Later research by Gosling et al. proposes a short ten-item version of the NEO-PI-R to capture the big five personality traits [16]. Although the precision of the NEO-PI-R cannot be reached with its shorter version, the so-called Ten Item Personality Measure (TIPI), the results can still precisely describe the personality of a subject. Just as the IUIPC privacy measures, the "big five" of personality can also be derived from blog or social network entries [7] of a user.

[2] Privacy Concern Scale

[3] Scale of Concern For Information Privacy

B. The influence of context factors on location privacy settings

According to a study by Benisch et al., there are several major context factors that influence people's decisions on location sharing: first the recipient of the location, the location itself, and the time of day and day of week the location is shared [2]. In contrast, earlier work identified mainly the requester's identity [11] as well as the user's activity/occasion [10] as important context factors. On social network sites, the post content or topic of the post [26] is often used as a criterion for the prediction, although the most recent work has shown that the post content only plays a minor role for predicting the privacy settings on social media sites using machine learning [15]. Interestingly, users are significantly more willing to share their location if they are paid for it. Even if it is clarified how the data is used and which negative consequences can arise from paid sharing, they do not change their decision [17]. Brush et al. [4] introduced different location obfuscation algorithms to users, and used a study to verify whether the users were able to understand the effects of the techniques and whether they were willing to use them in their location sharing applications. The authors were able to show both that the participants were able to understand the techniques, and also willing to use them.

Patil et al. [25] explored the user specification of location access rules by letting study participants define their access rules in free text form and later analyzing the most common context factors. According to their results, the privacy decision for sharing location settings is based on several factors, where the most significant are the recipients or requester of the location and the occasion or the position where the location is shared. Also, the granularity of the location plays an important role: There is a significant difference in whether the exact location is shared, or only the district or city where the person is located. In contrast to earlier work, the time and day of week play only a minor role in the participants' specifications. Also in the context of social network privacy, related work found the recipients of the post and the post topic to be important aspects of the sharing decision[26].

For our study, we oriented ourselves to the most recent results [25] and therefore used the occasion/topic of the location sharing and the receiving friend group as context factors. In addition to context factors, we also take the user's personality and privacy concerns into account (later called *individual features*), to deliver a privacy setting that is tailored for the user's needs. To the best of our knowledge, we are the first to allow the prediction of a fine-grained location sharing level (later called *privacy level*), that also makes it possible for example to display only the street or current city location, rather than simply disclosing or not disclosing an exact location.

C. Location privacy awareness and management tools

There are basically two different approaches to support users in defining their privacy rules: The first one tries to improve their overall understanding (*privacy awareness*) of their current privacy state and the set of rules, to help users to spot critical settings and to motivate them in specifying new rules to resolve the critical privacy settings. The second solution, in contrast, tries to automatically formulate rules for the user (*location management*), for example by using machine learning based on context factors or the user's personality. Some of them also use a combination of both approaches. One of these hybrid approaches is used in *Buddytracker*[18]: Whenever the current user location is requested, the user is notified by a notification bubble. He then has the possibility to allow or reject the inquiry. Studies have shown that after a training phase, users disclosed significantly fewer locations than before. Delphine et al. [8] published an approach that is tailored towards mobile information sharing, and allows one to set the privacy settings for location sharing, accelerometer, camera and microphone data using a radar-based user interface. Critical settings are identified and visualized to the user by concrete examples of possible consequences. Their study [9] has shown that users are significantly more willing to adapt their settings when using the UI. A similar approach by Tsai et al. [34] motivated users to restrict their privacy settings by a simple notification box, whenever a location was accessed.

Peoplefinder [32] is a tool that is more on the privacy management rather than the privacy awareness side: The tool maintains a friend list, where the user can add or remove friends who should be able to access his or her location. The access can be further restricted by specifying time-based access rules. Whenever a location access is granted, the user is notified by a small pop-up in the taskbar. In addition to manually specified rules, Peoplefinder also can predict an optimal set of rules using machine learning and a random forest classifier. Another privacy management tool by Johnson et al. offers three different functionalities for different use-cases: The *long-standing location sharing* always shares the current location, based on the geographical distance to the requestor. The longer the distance, the more abstractly the location is displayed to the requestor. The *proximity detection* functionality only displays which friends are in the direct surroundings of the user. Lastly, the *rendezvous* functionality shares the location only with a small set of friends at a user-defined time of the day, e.g. to meet up for partying on the weekend. Whereas the first two functionalities were evaluated as useless, the *rendezvous* functionality was perceived as very positive and useful by the study participants.

Research has also investigated how the location privacy for mobile smartphone apps can be improved: Fawaz et al. [14] published a tool that automatically decides which abstraction level is suitable for the apps installed on the smartphone's users, depending on the app functionality: Apps in the background or apps that have an advertising purpose are blocked from the location sensor, whereas apps requiring only a rough location estimation, like weather forecast apps, receive the city center instead of the exact GPS location. Lastly, *privacy shake* introduces a shake gesture to turn on or turn off location disclosure on the go. Although the idea seems promising, the gesture was not recognized well (recognition rate of less than 40%), which led to a major frustration for the study participants.

To conclude, the state of the art is to either try to improve privacy awareness and to motivate the user to restrict their privacy disclosure, or, like in our case, to try to automatically predict the privacy settings for the user. There are several approaches that use machine learning to decide to disclose or not disclose the location. But so far, to the best of our knowledge, there is no solution that generates individual privacy settings based on both context as well as individual factors, and that allows a fine-grained location-sharing policy, as recommended by the related literature [25], [26].

III. User study and correlation analysis

As described in the introduction, there is an effect called the *privacy paradox*[1], which means that the disclosement settings chosen on social network sites do not correspond to users' desired disclosement. We therefore decided against extracting the privacy settings from existing Facebook profiles for our study, and went for a questionnaire where users had to explicitly state their desired disclosement settings. In more detail, we used the following three-step approach in our study, which will be described in the next sections: First we gathered data for our *gold set* (the data set that contains both privacy/personality measures as well as location sharing privacy levels) using an online study. Using the *gold set*, we determined *whether* there are correlations between the personality/privacy measures and the location disclosure preferences in the second step. Finally, we discuss the observed relationships, and validate whether they can be leveraged for predicting the settings using regression algorithms. Related research [26], [25] uses different input features that we differentiate into *context features* and *individual features*. Whereas the *individual features* are different for every single user (like the personality or privacy measures), the *context features* do not depend on the user, but for example on the post (like the topic of the post). The authors have shown that there are notable differences ine sharing preferences on Facebook depending on personality and privacy attitudes (as individual features), as well as on the topic (as a context feature) of a post and the friend group that is receiving the post. We suspect that the same holds for shared location information, and will therefore investigate whether there is a significant influence on the location sharing setting. In more detail, we will use the friend groups and topics that have been found to be most frequently used in social media by related work [20], [26]. Furthermore, they propose a fine-grained sharing approach that offers more than just to *show* or *hide* the posts to a friend group. Their approach proposes five different privacy levels that also allow for a middle-of-the-road approach by, for example, showing only textual content while hiding photos or comments. In our study, we will also offer the users more than just to *show* or *hide* the GPS location. We also offer five intermediate sharing levels, which show only the current street, or the center of the city where the user is currently located. To minimize side effects and random variables as best as possible, we decided to capture the privacy and personality measures using a questionnaire, rather than extracting it out of the users' posts, although this has been shown to be possible[7]. The goal of our study is to find out *whether* it is possible to propose a privacy level based on the aforementioned input factors, rather than *how* an interface implementing our idea could look, and how such an interface might be perceived/accepted by a user. We therefore concentrated on collecting training data and measures about the precision, which can give us an insight on the feasability of the proposed approach, rather than evaluating the potential of a future UI implementation in terms of user acceptance and user experience by collecting qualitative data.

Privacy Level	**Displayed location**
1 - Exact Location	exact GPS location
2 - Street & city only	area of the whole street
3 - City only	city area
4 - Province only	area of the province
5 - Country only	area of the whole country
6 - Continent only	area of the continent
7 - No location	none

TABLE I. Privacy levels, their name and the description used in the online study.

A. Methodology

The study was conducted as an online survey using the software LimeSurvey.[4] 100 participants were recruited using Prolific Academic,[5] which allows us to select only participants that are actively using either a location sharing service or the "share location" functionality on Facebook/Google+ or Google Maps, for example.

Studies in the past have shown that participants who are recruited via online services, like in our case, lead to a similar quality of the results as when participants are recruited at a university [6]. The participants needed on average ten minutes to complete the questionnaire and were paid a compensation of £1 upon successful participation. To motivate the subjects to fill out the questionnaire honestly, the compensation was only paid after we checked the submitted data for plausibility. If the results from a subject were rejected, for example if he failed to answer the control questions correctly, a new participant was recruited to fill in the gap. Therefore we had exactly 100 results for the study.

The age of the participants ranged from 19 to 65 years (average 33.08, SD 9.14). The recruited audience was very diverse: We recruited students, self-employed workers, employees, and also homemakers.

The survey can be divided into two parts: In the first part, we posed the questions of the big five personal inventory, IUIPC and Westin questionnaires. The second part consisted of a table where the participant had to enter his or her location sharing preference in the form of a *privacy level* for each combination of friend group and topic of the post/occasion of the location sharing, resulting in $9 * 11 = 99$ individual privacy levels for each participant, 9900 in total. To ensure comparability of the results, we did not use example posts from the users' social network profiles, but used a general description like "Imagine you had to share a post about a family occasion with your friends from the sports club. How much detail about your location would you share within this combination". As privacy levels, we used the different location abstraction levels that are provided by the Google API, also described in Table I. We provided an explanation for each privacy level, topic and friend group at the bottom of the questionnaire page. The topics used in the study were the same as in related work[26], namely "family affairs", "events", "movies", "politics", "food", "work", "hobbies", "travel" and "sports". The study ended with a short text field to enter feedback or comments for the study. The procedure described in the last sections was reviewed and approved by the ethical review board of our institution.

[4] https://www.limesurvey.org, last accessed 09-07-2017

[5] https://www.prolific.ac/, last accessed 09-07-2017

IV. Results

Prior to the correlation analysis, we first analyzed whether the seven different privacy levels were used, or whether a binary decision is suitable enough for this purpose. All of the privacy levels were used by, at minimum, 18 of the 100 study participants. Interestingly, the most frequently used setting was not *exact location* or *no location*, as is offered by most service providers; it was *city only*, used by 93 participants. On average, the participants used it for 32.65% of all settings. *No location* was used second most frequently (22.53%) followed by *exact location* (17.84%) and *street only* (15.94%). Interestingly, the number of participants who used the *exact* or *street* level is higher (75 and 74, respectively) than the number choosing *no location* at least once. This leads to the assumption that users who tend to block a location for at least one friend group and topic tend to do so more often. *Province* and *Country* were used for around 5% of the settings, in total by 43% and 45% of the participants. The least frequently used privacy level is *Continent only*, used by only 18% of the participants, in total for 0.57% of all settings.

As stated in earlier sections, we are also interested in whether the topic and the receiving audience (friend group) influence the choice of location privacy level, and whether they should be included as *context features* in the prediction. The context features consist of categorical data that has been collected using repeated measures for the different topics or groups within the subjects. The suitable statistical test for this purpose is therefore a repeated measures ANOVA on the privacy levels with the group or topic as the inner subject factor. Before performing an ANOVA, we checked the data sets on sphericity using a Mauchly test. The Mauchly test was significant for both the groups comparison ($p < 0.001$) and the topics ($p < 0.001$). The results reported here are therefore the results of the Greenhouse-Geisser test, which is the suitable equivalent for non-spheric data. The statistical results for the different groups and topics indicate that both the topic and the audience *strongly* influence the decision. In contrast to earlier work that found the topic to be of minor influnce [15], the F-value for the topics ($F_{10,890} = 66.865, p < 0.001$) indicates an even stronger influence than the receiving group ($F_{8,1092} = 3.329, p = 0.001$).

In the next step, we tested whether the recorded *individual features*, e.g. the personality and privacy measures, also influence the sharing decision, and whether they should be included as input for a machine-learning based prediction. Both privacy and personality measures are ordinal data; we therefore performed a correlation analysis between the individual features and the privacy levels. As the shape of the data does not always provide a normal distribution, we decided to perform a Spearman correlation. The results are reported in Table II.

In total, we analyzed $n = 9900$ data sets. The largest correlation coefficients could be observed with the IUIPC privacy measures. Especially participants that have a high control and collection measure tend to prefer obfuscated locations more than others. The Westin privacy index has, concerning the privacy measures, the least correlation with the privacy levels. As earlier studies have already shown, the privacy index is too coarse-grained to facilitate predicting privacy decisions[35]. Also, the personality measures show a strong correlation, but in general with a smaller correlation coefficient than the IUIPC privacy measures. The more open, agreeable and extraverted a person is, the more she is willing to publish her exact location. In contrast, neuroticism and conscientiousness lead to a higher degree of location obfuscation.

Individual measure	rho	p
openness	-.071	<.001
extraversion	-.068	<.001
conscientousness	.060	<.001
agreeableness	-.024	.019
neuroticism	.037	<.001
collection	.167	<.001
control	.106	<.001
awareness	.050	<.001
privacy_index	.024	.018

TABLE II. Correlations between individual features and the privacy levels

A. Discussion of the results

The participants thoroughly used most of the offered privacy levels, indicating that there is a need for fine-grained location-sharing, other than just disclosing or not disclosing, as offered by most social networks or location sharing services. Although there are only a few settings that use the *continent only* privacy settings, nevertheless there is a notable number of users who would use this setting, if it were available. We therefore included *all* proposed privacy levels in the regression analysis. The results indicate that the *context features*, e.g. receiving group and topic/occasion of the location sharing, as well as the *individual features* like personality and privacy concerns, have an influence on the privacy level. The occasion of the location sharing has a notably larger influence on the (desired) privacy level than the receiving group. 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 [25]. The more the user wishes to have control over his own data and the amount of data that is collected, the more he tends to obfuscate his shared location. In contrast, extraverted and open-minded people tend to be more generous when it comes to location sharing.

Based on the previous results, we expect the following results for the regression analysis:

- Using only the topic/occasion already allows a first rough prediction of the privacy level
- The receiving group influences the regression result less than the occasion
- Privacy settings deliver better prediction results than the personality measures
- The regression algorithm performs notably better than an algorithm using only the mean value, ignoring the personality and privacy preferences of the user

V. Regression analysis

Although the *individual features* are ordinal-scaled, both structural features are plain categorical data. Categorical data has always been a problem when it comes to regression or machine learning, as these algorithms are based on scales and cannot use nominal data as an input. A frequently used and naive solution to this problem is the use of so-called *dummy variables* instead of a categorical variable: If the variable contains k categories, $k-1$ dummy variables are created to represent each category except for the last one. Only the dummy variable of the respective category of the categorical variable is set to 1; the others remain zero. If the categorical variable is set to category k, all dummy variables are set to zero. The regression is then started with the $k-1$ dummy variables as an input instead of the categorical variable. Although the technique is often used, it assumes an equal distance between all the categories, which is often not the case. Categorical regression (CATREG) [24] is a more sophisticated multivariate regression approach that transforms categorical variables into scales, using optimization algorithms to find a suitable order for the categories, and to determine the optimal scaling between them. It can also improve the prediction for ordinal measures or scales, if the equivalence of differences between the ordinal levels is uncertain. For our analyses, we use CATREG with a maximum of 100,000 steps and $\varepsilon = .00001$, meaning that if the optimization score increases less than ε after an optimization step, the process is ended prematurely. Topic and group have been entered as nominal variables and the individual features as ordinal values, as their origin is an ordinal Likert scale. We report for each combination of data sets the *adjusted coefficient of determination* (adjusted R^2) as well as the *apparent prediction error* (APE). The *apparent prediction error* (APE) is a value between zero and one denoting how well the prediction performs compared to a constant predictor that always takes the mean privacy level as the prediction. A value of zero means a perfect fit of the predicted value with the actual privacy level, whereas a value of one means the regression model is only as good as a constant prediction.

The adjusted R^2 is based on R^2, which denotes the *goodness of fit* of the regression model, indicating how well the regression line approximates the real data points. An R^2 of 1 indicates that the regression line perfectly fits the data, while a value of zero means that the model is no better than just taking the mean values [22]. To be more precise, R^2 is computed as the fraction of variance explained by the regression divided by the total variance.

It is hard to judge whether the current set of features is optimal by using the conventional R^2 score, as it always increases with an increasing number of features, even if the new feature does not notably increase the prediction precision. Thus a high R^2 does not mean that the optimal set of features is included; it can also be a result of overfitting. We therefore only report the *adjusted* R^2 in our results, to try to prevent this problem. The adjusted R^2 is always less than or equal to the R^2. Informally stated, when adding new features to the regression, *adjusted* R^2 only increases if a feature is added that also sufficiently increases the R^2 value; otherwise it decreases [22]. With an increasing number of features included in the regression, the penalty of adjusted R^2 for the new feature is also increasing. Therefore the *adjusted* R^2 can also be negative, especially if the model is overfitted.

The results in Table III show that regression with only the topic as an input already allows a prediction better ($APE = 0.966$) than the baseline (constantly predicting the mean value of all results), which is only slightly decreased when the receiving group is added as a feature ($APE = .964$). Introducing the personality measures notably reduces the APE ($APE = 0.900$) whereas adding the privacy measures to the structural features leads to a better regression model ($APE = .897$). If only the two feature sets with the best correlations (topic and IUIPC; see last section) are used, an apparent prediction error of .899 can be reached. Finally, taking all feature sets (structure, privacy and personality) into account leads to an APE of .808, meaning that the prediction is about 20% better than the constant predictor.

Input features	**adjusted R^2**	**APE**
Topic	3.3	.966
All structural	3.5	.964
Structural + personality	9.8	.900
Structural + IUIPC	10.1	.898
Structural + privacy	10.2	.897
Topic + IUIPC	10.0	.899
All	19.3	.808

TABLE III. COEFFICIENT OF DETERMINATION (R^2) AND APPARENT PREDICTION ERROR (APE) FROM THE REGRESSION USING CATREG FOR THE DIFFERENT SETS OF INPUT FEATURES.

VI. DISCUSSION AND FUTURE WORK

A. Precision of the prediction

As stated in the introduction, our goal was to achieve a prediction model which does not require knowledge of the user's location sharing behavior or privacy decisions in the past, allowing us to offer even new and unknown users a proposed location sharing setting. As described later in the discussion, user only need to either fill out a short non-technical questionnaire, or allow access to their social media profile for deriving the individual features. Inspired by the findings of related literature, we used both context factors, as well as user-specific *individual features* like privacy and personality desires. The receiving friend group and especially the topic or occasion of the location sharing already offer a first estimation of the privacy level. With additional personality or privacy features, the precision of the regression model can be further optimized, whereas *privacy* measures have a better effect on the precision than the personality. Considering both privacy and personality, as well as the structural features for the regression, allows us to increase the precision to about 20% better than the constant baseline. Whether this increase in prediction precision is suitable enough to assist users in their everyday social media usage has to be elaborated on in future work, using a fully implemented prototype. To conclude, we can say that although *structural features* are easier to collect, having *individual features* notably increases the precision of the model. As suspected by related literature, also when it comes to fine-grained location sharing, the Westin privacy index is too coarse-grained to allow a prediction of the privacy preferences [35]. Also, the group of recipients plays only a minor part, although it was found to be of some importance in related work [25].

B. The need for fine-grained privacy levels

Our results support the findings of related work [25], indicating that sharing of obfuscated locations is also needed: *All* of our privacy levels have been used by *some* users, although only 18% used the privacy level "Continent only" at least once. Interestingly, the most commonly used privacy level was "City only". To disclose the exact location, or none at all, was only chosen second and third most frequently, highlighting the need for a fine-grained location privacy system, which is so far only supported by some of the large social network or location service providers like Facebook or Google Maps. Whether the proposed seven location levels are optimal, especially whether the more rarely used options "Continent only" and "Country only" can be omitted, still has to be examined in future work.

C. Time needed to gather the individual measures

At first glance, the time saved by inferring the location sharing settings seems to be reduced by the fact that personality and/or privacy measures need to be recorded for a user, for example by filling out a questionnaire, before he can start to use the prediction. Research has also started finding a solution for reducing this additional user burden: Several related papers [7], [13] have shown that written text from either Facebook, Twitter or Youtube can be used to predict personality with a standard error of about 0.5 on a five-point scale. At the time of writing, we are about to conduct a study to observe whether privacy measures can also be predicted using posts from Twitter and Facebook. The results indicate that it is possible with a similar precision.

The additional user burden can therefore be significantly reduced by connecting to a Twitter or Facebook account, and using the post entries to automatically derive the needed privacy and personality measures. If the user still wants to invest the time, he can take the surveys to further increase the prediction precision.

D. Limitations

We used the data of 100 participants in our study, which delivered a total of 9900 data sets for the analyses. Although this data set size already allows a good estimate of the actual precision, we would still like to examine whether the precision can be further increased when the idea is implemented as a social network plugin or, in the best case, implemented by a social network provider, offering millions of data sets for the training and prediction of the privacy levels.

For our study, we used a finite set of location abstraction levels, topics and recipient groups (see section III-A for details). Although these item sets have been worked out in seperate user studies in related literature [26], [20], they cannot represent all possible topics and groups that might be of interest when sharing a location. For the study, we had to find a compromise between practicability and degree of realism, as it is not possible to present a trained system with individual topics and friend groups for each participant. A fully individualized system would operate better, but the results have shown that the precision can already be high with an unindividualized system.

Finally, it is also possible that a location belongs to more than one topic, or is at the boundary between topics. Although the evaluation assumes that always exactly one topic is selected, the prediction mechanism can also handle a post that is tagged by multiple topics. In this case, the prediction will give us several privacy policies, one for each topic. A merging algorithm would go through all of the user groups in each of the policies, and according to Ravichandran et al. [31], use for each observed group the according privacy level depending on the conservativeness ratio of the user in the merged policy. The same merging technique can be used for friends which appear in multiple friend groups. Although this merging is not yet implemented and evaluated, it will allow us to predict posts with an arbitrary number of topics in a future version.

E. Lessons learned

Fine-grained location sharing settings are becoming popular in social networks like Facebook or Google Maps. On one hand, it allows users to better protect their privacy, on the other hand it becomes even harder and more burdensome to correctly set the location sharing settings for all shared locations or posts.

We did not try to observe and predict actual privacy behavior, as it does not correspond to the real privacy desire, known as the privacy paradox[1]. Instead, we **explcitely asked the users for their desired location sharing level** in our study. We have shown that **there are significant correlations** between **structual features like the post topic or friend group** as well as **context factors like the user's personality or privacy preferences** and the desired location sharing level. In summary, we can say that the use of **structural features alone is not sufficient** for an acceptable prediction precision using a categorial regression. If either personality or privacy measures are available, either by filling out a questionnaire or through text feature extraction (see related work), these should be used for the prediction. If there is a **choice between privacy and personality measures**, the **privacy measures should be preferred**. To be more precise, the three measures of the 12-item IUIPC questionnaire are needed; the Westin privacy scale has only a minor effect. If the amount of input data should be reduced to a minimum, we recommend omitting the receiving group, the personality scale and the Westin privacy index, and only using the topic or occasion along with the IUIPC measures as an input. Using all context features togehter, we can achieve an **improvement of 20% correctly predicted settings using CATREG** compared to using a generic privacy settings template.

F. Future work

Our results indicate that the prediction of fine-grained privacy levels is possible using both context factors and individual personality and privacy measures. Although we collected 9900 data sets in the study, we are still interested in how far the precision can be increased with a larger data training set. For this purpose, we plan to implement the described approach as a social network plugin, and to collect a larger amount of data in an in-the-wild study.

Besides increasing the precision, we would also like to examine how our idea is perceived by the users, especially whether the precision of 20% above a constant prediction is sufficient to be accepted by users. We would like to implement our approach as a Facebook UI plugin, and conduct a user study with a fully functional prototype, to see how the UI and the prediction are perceived by users, and how often it is used to generate location sharing settings rather than doing this manually. The usage frequency over time, the usage frequency of the different location privacy levels, and the learning curve with our tool would be of special interest.

In our opinion, an automatic privacy setting derivation system can never substitute for the user. There is always the need for a second component that allows the user to have a clear and understandable overview of the privacy settings, as well as their consequences. The prediction algorithm can only support him in his work, by adjusting a major portion of the privacy settings. The user still has to do the fine-tuning in a UI, although the amount of work can be significantly reduced. Another aspect that we want to cover in a future user interface is the tradeoff between privacy and consequences, e.g. what consequences arise for both privacy and also the usefulness of the location sharing, when the user chooses a higher privacy level, including a higher obfuscation. Currently, the user does not get any hints about these factors when he chooses a location sharing level, for example on Facebook. A final implementation of our UI should include a detailed description on the privacy implications and effects for friends that can see the location, for each proposed setting.

VII. Conclusion

Location sharing has become more common over the last few years. Nevertheless, the privacy options provided by the location sharing providers are rather simplistic: to share or not to share. Whenever the user decides to limit the audience for the post, he has to set things manually. Research has therefore begun to automatically infer privacy rules based on context factors like occasion or time of day. Other researchers have shown that users prefer a fine-grained location disclosure functionality, including abstraction levels on the street or city level. Nevertheless, to the best of our knowledge, nobody has combined the influence of context factors *and* privacy and personality measures to perform a prediction of a fine-grained location privacy scale. We performed an online user study to first find out which context and individual factors are important, and how precise a prediction might be, using a categorical regression approach. Based on the results, we summarized some indications to give an idea of which data source should be used, based on the available data and willingness of the user to fill in additional questionnaires. The results indicate that the best results (nearly 20% better accuracy than a constant value) can be achieved using the occasion, personality and the IUIPC personality questionnaire, whereas a minimal set of occasion and IUIPC measures can also lead to a prediction precision of 10% above average. Although our work gives a first insight on which input features are important and how precisely a prediction can perform, we still have to examine whether the precision is sufficient to implement an acceptable privacy setting prediction tool on a social network. Furthermore, we would still like to integrate our approach into a large-scale social network or location sharing service, to check whether the prediction precision can be increased with a large data set, and how well the idea is perceived and accepted by users.

References

[1] S. B. Barnes, "A privacy paradox: Social networking in the united states." *First Monday*, vol. 11, no. 9, 2006.

[2] M. Benisch, P. G. Kelley, N. Sadeh, and L. F. Cranor, "Capturing location-privacy preferences: Quantifying accuracy and user-burden tradeoffs," *Personal Ubiquitous Comput.*, vol. 15, no. 7, pp. 679–694, Oct. 2011. [Online]. Available: http://dx.doi.org/10.1007/s00779-010-0346-0

[3] J. Block, "A Contrarian View of the Five-Factor Approach to Personality Description," *Psychological Bulletin*, vol. 117, pp. 187–215, 1995.

[4] A. B. Brush, J. Krumm, and J. Scott, "Exploring end user preferences for location obfuscation, location-based services, and the value of location," in *Proceedings of the 12th ACM International Conference on Ubiquitous Computing*, ser. UbiComp '10. New York, NY, USA: ACM, 2010, pp. 95–104. [Online]. Available: http://doi.acm.org/10.1145/1864349.1864381

[5] T. Buchanan, C. Paine, A. N. Joinson, and U.-D. Reips, "Development of measures of online privacy concern and protection for use on the internet," *Journal of the American Society for Information Science and Technology*, vol. 58, no. 2, pp. 157–165, 2007. [Online]. Available: http://dx.doi.org/10.1002/asi.20459

[6] M. Buhrmester, T. Kwang, and S. Gosling, "Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data?" *Perspectives on Psychological Science*, vol. 6, no. 1, pp. 3–5, 2011.

[7] J. Chen, E. Haber, R. Kang, G. Hsieh, and J. Mahmud, "Making use of derived personality: The case of social media ad targeting," in *International AAAI Conference on Web and Social Media*, 2015. [Online]. Available: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM15/paper/view/10508

[8] D. Christin, M. Michalak, and M. Hollick, "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*, ser. MoMM '13. New York, NY, USA: ACM, 2013, pp. 445:445–445:454. [Online]. Available: http://doi.acm.org/10.1145/2536853.2536861

[9] D. Christin, A. Reinhardt, M. Hollick, and K. Trumpold, "Exploring user preferences for privacy interfaces in mobile sensing applications," in *Proceedings of the 11th International Conference on Mobile and Ubiquitous Multimedia*, ser. MUM '12. New York, NY, USA: ACM, 2012, pp. 14:1–14:10. [Online]. Available: http://doi.acm.org/10.1145/2406367.2406385

[10] K. Connelly, A. Khalil, and Y. Liu, "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*, ser. INTERACT'07. Berlin, Heidelberg: Springer-Verlag, 2007, pp. 620–623. [Online]. Available: http://dl.acm.org/citation.cfm?id=1776994.1777074

[11] S. Consolvo, I. E. Smith, T. Matthews, A. LaMarca, J. Tabert, and P. Powledge, "Location disclosure to social relations: Why, when, & what people want to share," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '05. New York, NY, USA: ACM, 2005, pp. 81–90. [Online]. Available: http://doi.acm.org/10.1145/1054972.1054985

[12] P. Costa, R. McCrae, and I. Psychological Assessment Resources, *Revised NEO Personality Inventory (NEO PI-R) and NEO Five-Factor Inventory (NEO-FFI)*. Psychological Assessment Resources, 1992. [Online]. Available: https://books.google.de/books?id=mp3zNwAACAAJ

[13] G. Farnadi, G. Sitaraman, S. Sushmita, F. Celli, M. Kosinski, D. Stillwell, S. Davalos, M.-F. Moens, and M. De Cock, "Computational personality recognition in social media," *User Modeling and User-Adapted Interaction*, vol. 26, no. 2, pp. 109–142, Jun 2016. [Online]. Available: https://doi.org/10.1007/s11257-016-9171-0

[14] K. Fawaz and K. G. Shin, "Location privacy protection for smartphone users," in *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '14. New York, NY, USA: ACM, 2014, pp. 239–250. [Online]. Available: http://doi.acm.org/10.1145/2660267.2660270

[15] C. Fiesler, M. Dye, J. L. Feuston, C. Hiruncharoenvate, C. Hutto, S. Morrison, P. Khanipour Roshan, U. Pavalanathan, A. S. Bruckman, M. De Choudhury, and E. Gilbert, "What (or who) is public?: Privacy settings and social media content sharing," in *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, ser. CSCW '17. New York, NY, USA: ACM, 2017, pp. 567–580. [Online]. Available: http://doi.acm.org/10.1145/2998181.2998223

[16] S. D. Gosling, P. J. Rentfrow, and W. B. Swann, "A Very Brief Measure of the Big-Five Personality Domains," *Journal of Research in Personality*, vol. 37, no. 6, pp. 504–528, December 2003. [Online]. Available: http://dx.doi.org/10.1016/S0092-6566(03)00046-1

[17] L. Hutton, T. Henderson, and A. Kapadia, “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. [Online]. Available: http://doi.acm.org/10.1145/2627393.2627416

[18] L. Jedrzejczyk, B. A. Price, A. K. Bandara, and B. Nuseibeh, “On the impact of real-time feedback on users’ behaviour in mobile location-sharing applications,” in *SOUPS ’10: Proceedings of the Sixth Symposium on Usable Privacy and Security*. New York, NY, USA: ACM, July 2010, pp. 1–12. [Online]. Available: http://oro.open.ac.uk/22571/

[19] O. P. John and S. Srivastava, “The Big Five Trait Taxonomy: History, Measurement, and Theoretical Perspectives,” in *Handbook of Personality: Theory and Research*, 2nd ed., L. A. Pervin and O. P. John, Eds. New York: Guilford Press, 1999, pp. 102–138. [Online]. Available: http://darkwing.uoregon.edu/~{}sanjay/pubs/bigfive.pdf

[20] P. G. Kelley, R. Brewer, Y. Mayer, L. F. Cranor, and N. Sadeh, *An Investigation into Facebook Friend Grouping*, ser. INTERACT’11. Berlin, Heidelberg: Springer-Verlag, 2011, pp. 216–233. [Online]. Available: http://dl.acm.org/citation.cfm?id=2042182.2042202

[21] P. Kumaraguru and L. F. Cranor, “Privacy indexes: A survey of westin’s studies,” *ISRI Technical Report*, 2005.

[22] M. Kutner, *Applied Linear Statistical Models*, ser. McGraw-Hill international edition. McGraw-Hill Irwin, 2005. [Online]. Available: https://books.google.de/books?id=0xqCAAAACAAJ

[23] N. K. Malhotra, S. S. Kim, and J. Agarwal, “Internet users’ information privacy concerns (iuipc): The construct, the scale, and a causal model,” *Info. Sys. Research*, vol. 15, no. 4, pp. 336–355, Dec. 2004. [Online]. Available: http://dx.doi.org/10.1287/isre.1040.0032

[24] J. J. Meulman, “Prediction and classification in nonlinear data analysis: Something old, something new, something borrowed, something blue,” *Psychometrika*, vol. 68, no. 4, pp. 493–517, Dec 2003. [Online]. Available: https://doi.org/10.1007/BF02295607

[25] S. Patil, Y. Le Gall, A. J. Lee, and A. Kapadia, *My Privacy Policy: Exploring End-user Specification of Free-form Location Access Rules*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 86–97. [Online]. Available: https://doi.org/10.1007/978-3-642-34638-5_8

[26] F. Raber, F. Kosmalla, T. Schneeberger, and A. Krueger, “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*, R. Bernhaupt, G.Dalvi, A. Joshi, J. O’Neill, and M. Winckler, Eds., IFIP. Springer, 2017.

[27] F. Raber and A. Krueger, “Towards understanding the infuence 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*, R. Bernhaupt, G.Dalvi, A. Joshi, J. O’Neill, and M. Winckler, Eds., IFIP. Springer, 2017.

[28] F. Raber and A. Krüger, “Privacy perceiver: Using social network posts to derive users’ privacy measures,” in *Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization*, ser. UMAP ’18. New York, NY, USA: ACM, 2018, pp. 227–232. [Online]. Available: http://doi.acm.org/10.1145/3213586.3225228

[29] F. Raber and N. Vossebein, “Uretail: Privacy user interfaces for intelligent retail stores,” in *16th IFIP TC 13 International Conference on Human-Computer Interaction — INTERACT 2017 - Volume 10516*. Berlin, Heidelberg: Springer-Verlag, 2017, pp. 473–477. [Online]. Available: https://doi.org/10.1007/978-3-319-68059-0_54

[30] F. Raber, D. Ziemann, and A. Krüger, “The ’retailio’ privacy wizard: Assisting users with privacy settings for intelligent retail stores,” in *EuroUSEC ’18 : 3rd European Workshop on Usable Security. EuroUSEC European Workshop on Usable Security (EuroUSEC-18), 3rd, located at IEEE Conference on Security & Privacy, April 23, London, UCL, United Kingdom*, C. Weir and M. Mazurek, Eds. Internet Society, 2018.

[31] R. Ravichandran, M. Benisch, P. G. Kelley, and N. Sadeh, “Capturing social networking privacy preferences,” *Soups*, p. 1, 2009. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1572532.1572587

[32] N. Sadeh, J. Hong, L. Cranor, I. Fette, P. Kelley, M. Prabaker, and J. Rao, “Understanding and capturing people’s privacy policies in a mobile social networking application,” *Personal Ubiquitous Comput.*, vol. 13, no. 6, pp. 401–412, Aug. 2009. [Online]. Available: http://dx.doi.org/10.1007/s00779-008-0214-3

[33] H. J. Smith and S. J. Milberg, “Information privacy: Measuring individuals’ concerns about organizational practices,” *MIS Q.*, vol. 20, no. 2, pp. 167–196, Jun. 1996. [Online]. Available: http://dx.doi.org/10.2307/249477

[34] J. Y. Tsai, P. Kelley, P. Drielsma, L. F. Cranor, J. Hong, and N. Sadeh, “Who’s viewed you?: The impact of feedback in a mobile location-sharing application,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI ’09. New York, NY, USA: ACM, 2009, pp. 2003–2012. [Online]. Available: http://doi.acm.org/10.1145/1518701.1519005

[35] A. Woodruff, V. Pihur, A. Acquisti, S. Consolvo, L. Schmidt, and L. Brandimarte, “Would a privacy fundamentalist sell their dna for $1000... if nothing bad happened thereafter? a study of the westin categories, behavior intentions, and consequences,” in *Proceedings of the Tenth Symposium on Usable Privacy and Security (SOUPS)*, ACM. New York, NY: ACM, 2014, iAPP SOUPS Privacy Award Winner. [Online]. Available: https://www.usenix.org/conference/soups2014/proceedings/presentation/woodruff