Differences in Smartphone Usage: Validating, Evaluating, and Predicting Mobile User Intimacy

Mattia Gustarini^a, Marcello Paolo Scipioni^b, Marios Fanourakis^a, Katarzyna Wac^{a,c}

- a. University of Geneva (UNIGE), Informatics University Center, Information Science Institute, 1227 Carouge, Geneva, Switzerland, {firstname.lastname}@unige.ch
- b. Università della Svizzera Italiana (USI), Faculty of Informatics, Ubiquitous Computing, 6904 Lugano, Ticino, Switzerland, marcello.paolo.scipioni@usi.ch
- c. University of Copenhagen (UCPH), Department of Computer Science, 2100 København, Denmark, wac@di.ku.dk

Abstract

We analyze the users' intimacy to investigate the differences in smartphone usage, considering the user's location and number and kind of people physically around the user. With a first user study we (1) validate the intimacy concept, (2) evaluate its correlation to smartphone usage features and (3) we computationally model it. Shorter, more frequent, and less engaging interactions take place when intimacy is lower, while longer, less frequent, and engaging interactions when intimacy is higher. With a second user study, we investigate the intimacy predictability in practice. Location-time features are predictive for the intimacy, and other smartphone-based features can improve the intimacy prediction accuracy.

Keywords: user smartphone engagement; user context; user location; people around user; RPC ranking classification

1 Introduction

Smartphones follow us everywhere in our everyday life. People have become accustomed to using smartphones at home as well as at work, in public transportation or while walking in the city, sitting in a pub or on the go, with the family or alone, on vacation, and in many more situations. A recent study evaluated that people use their smartphones more than 200 times a day [1]; however people do not use

their smartphones the same way in every situation. While sometimes a quick glance at the phone is enough to check notifications status, in other cases people engage in longer sessions, *e.g.*, to answer emails, buy food, browse Facebook, search for a place on a map, or check the balance on the credit card [2].

In this paper, we first identified and fused together three objective context elements into a new subjective context variable, that we called *intimacy*. The three components are the place where a user is (*e.g.* at home, at work, or on the bus), the number of people around the user, and the kind of people (*e.g.* friends, family or strangers). We validate the intimacy variable by analyzing how the users' (subjective) perception of this intimacy varies when they are in different situations defined by (objective) context elements like location, the number and kind of people.

In this first step, we designed a one-month user study with 42 participants. In this study, users were asked to self-report through Experience Sampling Method (ESM) both their objective context, *i.e.*, location, number and kind of people, and subjective context, *i.e.*, their intimacy and mood (expressed by 'arousal' and 'valence'). We confirmed that the three context elements, location and number and kind of people, are significantly related to intimacy and that users use the smartphone differently depending on their perception of intimacy. We also show that the intimacy significantly relates to the 'valence' component of the user's mood.

Second, in addition to the self-reported elements collected through the ESM questionnaire, we automatically gathered smartphone usage data through a logger that was running in the background on each of the participants' phones. We evaluated the relation between intimacy and mobile device interaction features derived from the automatically logged phone usage data. Our findings highlight that the perceived intimacy of users correlates significantly with the duration of users' interaction with their smartphones and with the kind of users' interaction, *i.e.* whether users only glance at their phone without launching any applications, or whether they interact with several applications. In particular, users perform shorter, more frequent, and less engaging tasks when their intimacy is lower, and longer, less frequent, engaging interactions when their intimacy is higher.

Finally, we show that it is feasible to create a low-complexity computational model of intimacy state enabling to predict this state for mobile users. Furthermore, we have created several intimacy models based mainly on the users' location features. We performed two experiments to select the most predictive

features that we then used in a second user study with 31 participants over a period of four weeks. During this study we automatically predicted the intimacy state of mobile users and tested it against the ground truth collected from the users, to evaluate the accuracy of our model *in the wild* and find its computational limitations. Additionally, we collected new features potentially enabling us to increase the precision of the predictions in the future.

Based on the conducted research we conclude that the *intimacy* concept is a suitable candidate for further refining certain mobile application design aspects, such as refining the design of interfaces and patterns of interaction, or alleviating intrusiveness (*e.g.*, in case there are too many notifications) to evaluate real user needs. Additionally, we proved that it is feasible to predict intimacy using the users' smartphones starting with location-based features and that there is room to increase further the model accuracy with new variables.

2 Defining Intimacy

As reported by Prager [3], the concept of 'intimacy' is complex, and it has several definitions. The word 'intimacy', from the Latin 'intimare', means 'impress, make familiar,' (from 'intimus,' *i.e.* 'inmost') [4]. According to the dictionary, the adjective 'intimate' describes the context (place or setting) as familiar, being private and comfortable [5]. In our study, we refer to intimacy as a location/place or an environmental situation ("a cozy and private or relaxed atmosphere" [4]).

The location is one of the elements connected with the perception of intimacy. Manzo shows [6] how being at home or in another location can influence several traits of the people personality, such as emotions and perception of the environment. Seamon and Sowers build upon that by researching report on the work of Relph [7] researching about the deep meaning of people's attachment to a location/place. Relph [7] indicates 'home' as the most intimate place. Thus, we consider the location/place as a potentially predictive variable for the individual's intimacy.

Besides the place, another component that influences intimacy and describes the current setting is the presence of other people and the relationship of a person with them. Saegert [8] explains how the environment of people influences their cognitive perception of it. The more people are physically around, the more complicated the situation becomes cognitively, and the person tends to be distracted more. Diener and Seligman [9] point out how the presence of closest people contributes to the well-being of

people. Miller and Lefcourt [10] consider these relations as social intimacy. In their study, they interviewed undergraduate students about their relationships with friends, acquaintances, and family members to define the characteristics of the relationships that they consider intimate.

Based on the previous work, we considered (1) location, (2) number and (3) kind of people physically around the user as the three top contextual elements related to the users' perception of intimacy.

We are not investigating the relation of intimacy with other concepts such as love, sexuality, self-disclosure, or privacy [11], [12]. While there is a relation between intimacy and privacy [13], we only refer to intimacy as a perception of *attachment to and familiarity with* the three context elements: location, number and kind of people analyzed in our study. Such perception does not aim at analyzing communication with others or controlling the disclosure of information about oneself, or at capturing the quality of users' relationships as defined, for example, by Westin [12], but focuses solely on a subjective evaluation of one's current situation.

Based on this interpretation of intimacy, we hypothesize that when people are at home having a dinner with their families discussing family matters, they feel intimate, while when they are on a bus with strangers, they do not feel intimate.

3 Related Work

We based the intimacy concept on the elements belonging to the user's context. Past and ongoing research have focused on the understanding of the user's context, *i.e.*, what are the context elements, context modeling methods, context discovery, and so on. Comprehensive survey papers on context are those by Chen and Kotz [14], Bolchini *et al.* [15], Bettini *et al.* [16], Ye *et al.* [17], enumerating the existing context definitions, *e.g.*, the one of A. K. Dey [18], and investigating in depth their field of interest, *e.g.*, application design, situation identification, and context modeling.

Many works focused on understanding the context of users and how this influences the use of (mobile) devices. Eagle and Pentland [19], [20] used mobile phone Bluetooth capability to identify the users' social context and study how they were accessing and consuming information in different social situations. Adams *et al.* [21] focused on the users' colocation to establish social ties, using mobile phone GPS and Bluetooth. They analyzed how social interactions were influencing the use of two applications: a social photo and video album and a blog auto-generating content from social information. Zheng and Ni

[22] presented an innovative system to detect social circles using smartphone Bluetooth and envisioned the use of these circles in combination with social network apps usage.

Researchers also investigated the subjective nature of smartphone usage. Falaki *et al.* [23] showed how individual usage patterns can be exploited to personalize battery consumption prediction. Soikkeli *et al.* [24] focused on how smartphone usage changes depending on location. Differently from this work, we focus on smartphone usage patterns in a wider domain than the location, also considering the number and kind of people surrounding the user. Additionally, we propose to analyze the smartphone use depending on a subjective context variable, *i.e.*, the intimacy. Banovic *et al.* [2] researched how long users are interacting with smartphones by labeling the smartphone usage sessions in three categories: ‘glance’, ‘review’, and ‘engage’. They proved how these categories of interactions were able to represent the users’ general behavior while emphasizing that each user keeps her/his unique interaction pattern. We build upon this research by employing the proposed categorization of interaction sessions to carry out our analysis of smartphone usage.

Finally, researchers also considered the importance of subjective context elements, for instance by detecting users’ mood [25] or stress [26] states, by analyzing smartphone usage. In our initial research, Gustarini and Wac [27], we have already investigated how intimacy influences some aspects of smartphone usage regarding screen touches and applications used. Particularly, in [27] we investigated how intimacy influences the use of mobile applications at the level of application categories (which are the categories of apps that users use in low/high intimacy), with specific emphasis on the applications belonging to the social-communication category specifically (*e.g.*, text or call). The main contributions of our past work [27] are as follows. (1) When users switch between many applications in a single smartphone usage session, they are more likely to be intimate. (2) When users use some applications actively for a long time (*i.e.*, we noted long interactions and many screen touches) they are more likely to be intimate. (3) Users use messaging applications mostly when intimate and prefer to browse the internet when in lower intimacy. However in [27], we did not provide a complete definition of intimacy, and we did not verify our assumptions about its relation to location and number and kind of people around the users. Additionally in this work, we evaluate the smartphone interaction duration, the number of interactions, and the type of interaction (in several fixed time windows) and we provide a first version of the intimacy prediction model.

To our knowledge, no previous work has investigated users' perception of *intimacy* as we define it in this paper, nor described high-level smartphone usage patterns changes depending on this subjective context variable, and finally, no one ever attempted to model and predict intimacy as defined here.

4 Experimental Methods

4.1 User Study

We conducted a user study, for which we recruited 42 people (18 female and 24 male), aged 18 to 45 years, from a city in the US (22) and a city in Switzerland (20). The recruitment was performed using online announcements in Craigslist in the US and from our living lab participants in Switzerland. An online entry survey was performed to assess the participant ownership of a smartphone with an experience of at least six months. We involved the participants in the study for a minimum of 22 days to a maximum of 36 (average of 27 days). They were students or employees in different working fields, using their personal Android OS-based smartphones.

4.2 Data Collection

4.2.1 Experience Sampling Method

We required the study participants to answer an Experience Sampling Method (ESM) based survey about their intimacy appearing on their smartphone screen. Each ESM survey - a so-called "*beep*" - appeared randomly, but uniformly, eight times per day, between 8 am and 11 pm. Additional *beeps* showed each time the participants plugged their smartphone for charging. The application recorded the answers and sent them over the Internet to our dedicated server. Each *beep* had several questions about the participants' current context and feelings. (1) We asked about their subjective perception of *intimacy* via a six points Likert scale (following the answers entries order, from 1 'completely' intimate - to 6 'not at all' intimate); (2) the number of people around the participant ('alone', '1', '2-10', '11-20', '21-40', '40+'); (3) the kind of people around ('co-workers/classmates', 'family', 'friends', 'girl/boyfriend', 'other', 'strangers'); (4) their current semantic location ('bus', 'home', 'other', 'pub', 'school', 'shopping center', 'street', 'work'); (5 and 6) users indicated their emotional state using the self-assessment manikin (SAM)

scale [28], with two dimensions: valence and arousal (scaled from -4 to 4). Valence options range from ‘unpleasant’ to ‘pleasant’ and arousal from ‘calm’ to ‘highly activated’.

We informed participants about the definition of intimacy at the study entry. We instructed them to indicate their perception of intimacy in their current situation, *i.e.* the location, number and kind of people around them when answering the survey. Each answer is independent of others. This approach creates an annotated diary of the users’ context minimizing the recall error, *i.e.*, estimates of the user’s current context randomly but uniformly during the day [29].

4.2.2 Smartphone Logged Data

Besides issuing the ESM questionnaire, our research application was logging several system- and user-generated events as they occurred. We have captured the smartphone screen events, finger touches, and the apps that participants were using. From these logged events we elaborate smartphone usage variables. Each minute, we also captured the unique identification number of an access point to a cellular network (CellID) to which users’ smartphones were connected. We used them to define locations subsequently employed in the intimacy model. We provide more details in Section 7.1.1.

5 Validating the Intimacy Definition

To ensure that the intimacy scores we collected via ESMs were accurately expressing the intimacy of users, we checked if certain direct relations were in place. For instance, one would expect that in more familiar places, such as home, and with more familiar people, users would perceive higher intimacy (more attachment to these context elements) than in unfamiliar places or with less familiar people. Therefore, we evaluated how the variation of the (subjective) intimacy scores collected through the ESM questionnaires bounds to the variation of the collected (objective) context features.

5.1 General Statistics on ESM Surveys

The application issued a total of 12497 ESM surveys (on average ~11 per user, per day) and we collected 6509 answers from all the users (~5.7 per user, per day). In Figure 1 we present how each user contributed to the answers. Apart from a few cases, most of the participants reached at least 100 answers (~3.7 answers per day) with some users reaching over 200. Additionally, in Figure 2 we present the mean intimacy score for all the participants over all their intimacy ESMs.

5.2 ESM Surveys Data Preparation

Our data has two particularities that we need to take into account for a reliable analysis of the data. (1) As we show in Figure 2, users have a different overall perception of intimacy and tend to employ different subjective scales of intimacy; hence to be able to compare scores across the users we need to normalize these scales across all users. (2) As we present in Figure 1, users contribute differently to the dataset. For example, user 21 is a significant contributor, as well as participants 17, 18, 19 and 33. If we do not properly prepare the data for analysis, the users providing this information may bias the results.

Following the method by Larson and Delespaul [30] we transformed the ESM data before the hypothesis-testing step.

Firstly, to address the individual intimacy scales problem, we standardized the users' intimacy score using z-scores [30]. To do so, we scaled the intimacy score for each user by centering her/his score *mean* to 0; representing his/her 'usual' score. This scaling results in negative values when the user is in a higher intimacy than usual, and in positive values for a lower perceived intimacy.

Secondly, we aggregated the ESM answers of each user transforming the *beeps file* into a *subjects file*. In the *beeps file*, each row represents the answers a user gave for each ESM (*i.e.*, the answers to the six questions). This file contains as many rows as the total of all the surveys answered. Instead, in the *subjects file*, we aggregated the intimacy value as the mean value over all the different situations encountered by the users (*i.e.*, the combinations of the answers to the ESM questions 'where', 'number of people', 'kind of people'). For example, as a situation, the user was on the street ('where'), with 2-10 persons ('number of people'), being strangers ('kind of people'). We took all of her/his answers ('beeps') in this situation and aggregated the single intimacy scores with its mean. This *subjects file* contains as many rows as the number of different situations encountered by each user for all the users.

Differently from the beep-level analysis, by using the subject as the main unit of analysis, the assumption of independence between samples is not violated, and we avoid issues such as an inflated number of samples (*i.e.*, when a small number of individuals produce the majority of samples). For example, in our case we avoid that user 21, being a significant contributor of the data, defines a trend in the results (Figure 1). The disadvantage of using this approach is that we may not capture all the true relationships in our data [30].

5.3 Intimacy vs. Place, Kind and Number of People Around the User

In this section, we present how the intimacy concept correlates with place, number and kind of people around the user. Usually, ANOVA is used for such analysis. However, because of the nature of the ESM method, the questionnaire appeared randomly but uniformly during the day, there is no guarantee that each user happens to be in all possible situations (Lathia *et al.* [31]). Given the unbalanced nature of the data, *i.e.*, not all the situations are distributed equally and not all users fully covered them, we performed a Linear Mixed Model (LMM) analysis [32]. Differently from ANOVA, the LMM analysis can deal with missing data without the need to discard the incomplete features and thus losing information. We performed LMM with users as subjects and the context intimacy z-score as a dependent variable. We modeled the effect of the three contextual elements as factors ('where', 'number' and 'kind of people') with a random intercept to take into account the variance of intimacy between users. Finally, for all the three context elements we tested which of their possible values had significant different mean intimacy. For this task, we conducted pairwise post hoc tests with Bonferroni correction.

5.3.1 Semantic Place ('where')

With the result of the LMM analysis, we indicate that the place has a significant effect on the subjective perception of intimacy, $F(7, 1117) = 16.674, p < 0.001$. In the ESM survey, users could specify one of the following eight categories of place: 'bus,' 'home,' 'other' (all the rarely selected places), 'pub,' 'school,' 'shopping center,' 'street,' 'work.' Not all the users have been in all 8 places ('bus' = 19 participants, 'home' = 42, 'other' = 42, 'pub' = 30, 'school' = 31, 'shopping center' = 28, 'street' = 38, and 'work' = 37).

With pairwise comparison post hoc tests, we show that users perceive more important places, like 'home,' as more intimate (more familiar) than less visited ones. The place 'home' (Figure 3, A1) is significantly different from all other places belonging to group A2 $p < 0.001$. The members of group B1, 'Work' $p = 0.002$, 'school' $p = 0.034$, and 'other' $p = 0.006$ are more intimate than 'street' (B2). All other places differ amongst them, but not significantly.

The model estimated a z-score intimacy mean for 'home' of -0.02. This value is very close to 0 - representing the scaled average value of intimacy for all the users. An interpretation of this result can indicate that participants being at 'home' have their mean value level of intimacy perception, and

elsewhere people feel less intimate than home (*i.e.*, increasing z-score). The intimacy perception seems to relate to the time that an individual is used to spend in a given place. Since most of our users spent most of the time at ‘home’, it is reasonable to notice that it has higher familiarity and attachment. Then, places like ‘bus’, ‘shopping center’, and ‘street’ are visited quickly and occasionally (users perceive lower intimacy and thus reduced familiarity and attachment).

5.3.2 Number of People Around (‘number of people’)

With the LMM analysis result for this context element we indicate that the number of people has a significant effect on intimacy, $F(4, 1117) = 20.940, p < 0.001$. In the ESM survey, users could specify how many people were surrounding them by selecting one of the following six categories: ‘alone,’ ‘1,’ ‘2-10,’ ‘11-20,’ ‘21-40,’ ‘40+’ people. For some users not all the categories are present, in particular in the categories with many people around (‘alone’ = 41 users, ‘1’ = 42, ‘2-10’ = 42, ‘11-20’ = 41, ‘21-40’ = 37, ‘40+’ = 32).

With post hoc tests we show that the fewer number of people around the user, the higher is the users’ perceived intimacy. The category ‘0’ (‘alone’, Figure 4, A1) is significantly different from ‘1’ $p = 0.007$ (A2), and from the others $p < 0.001$ (A3 and A4). The category ‘1’ (A2) is significantly different from: ‘2-10’ $p = 0.001$ (A3), and ‘11-20’, ‘21-40’, and ‘40+’ $p < 0.001$ (A4). The category ‘2-10’ people (A3) is also significantly different from all the members of group A4 with $p < 0.001$. Finally, ‘11-20’, ‘21-40’, and ‘40+’ people (A4) are not significantly different from each other, but different from all the members of groups A1, A2, and A3 with $p < 0.001$.

In particular, users perceive differently being alone or with one person (more familiar) than being with 2-10 people and with 20+. Most probably, the ‘2-10’ category contains the threshold for which the number of people changes the perception from a high-perceived intimacy (higher familiarity and attachment) to a lower one. Having more than 20+ people around a user, results in low intimacy; for any higher number of people the intimacy level remains constantly low.

5.3.3 Kind of People Around (‘kind of people’)

The LMM results show that the factor ‘kind of people’ has a significant effect on the context intimacy mean, $F(5, 1117) = 17.450, p < 0.001$. In the ESM survey, users could specify what kind of people were surrounding them by selecting one of the following six categories: ‘co-workers/classmates,’ ‘family,’

‘friends,’ ‘girl/boyfriend,’ ‘other,’ ‘strangers.’ Also in this case, not all the users choose all the available categories (‘co-workers/classmates’ = 37, ‘family’ = 30, ‘friends’ = 33, ‘girl/boyfriend’ = 17, ‘other’ = 41, representing other possible combinations rarely rated, and ‘strangers’ = 41).

With post hoc tests, we could identify two main groups of ‘kind of people’. One accounting for people closer to the user (‘family’, ‘girl/boyfriend’, and ‘friends’ - Figure 5, A1), and another one with people that are less closely related (‘co-workers/classmates’ and ‘strangers’ - A2). The former contains the categories with which the users perceive high intimacy while the latter group contains the categories associated with lower perceived intimacy. The category ‘other’ does not belong to any of these two groups, probably due to its heterogeneous composition. The categories belonging to the group of closest people (A1) are all significantly different from the not closest one $p < 0.001$ (A2). ‘Family’ (B1) is also significantly different from ‘other’ $p = 0.046$ (B2), and there is a significant difference between ‘other’ (C1) and ‘stranger’ $p < 0.001$ (C2).

6 Relation of Intimacy with Mood and Smartphone Usage

This section provides answers to the questions: does intimacy relate to trending (*i.e.*, currently studied) subjective context like mood? Moreover, does intimacy relate to smartphone usage?

6.1 Intimacy and Mood Components: Valence and Arousal

Since we collected the mood components ‘valence’ and ‘arousal’ using ESM surveys, we prepared the data and analyzed it in the same way as we did in the previous section, *i.e.* we performed an LMM analysis and post hoc tests with Bonferroni correction.

Using the self-assessment manikin (SAM) scale [28] users could specify their subjective value of ‘valence’ from ‘-4 - unpleasant’ to ‘4 - pleasant’ with the neutral value as ‘0’. Also in this case, not all users selected all the categories (‘-4’ = 12, ‘-3’ = 15, ‘-2’ = 22, ‘-1’ = 28, ‘0’ = 38, ‘1’ = 34, ‘2’ = 37, ‘3’ = 37, ‘4’ = 28).

Valence has a significant relation with intimacy, $F(8, 242) = 3.045$, $p = 0.003$. With post hoc tests we show a clear pattern (Figure 6): when users are in higher intimacy (low-negative z-score), they perceive the situation as more pleasant than when they are in lower intimacy (high-positive z-scores). We show that intimacy has a relation with valence that can be leveraged further, *i.e.*, to work with mood related

tasks. After comparing the mean values of intimacy among all valence values, the only significant difference we found is between the less pleasant state of ‘-1’ and the more pleasant states ‘2’ $p = 0.039$ and ‘3’ $p = 0.010$. The rest of the comparisons did not lead to significant differences (all $p > 0.05$). This result is probably due to a high number of valence categories, 9, which would have required more ESM survey answers.

We also investigated the relation between intimacy and ‘arousal’, but we could not find any significant relation, $F(8, 244) = 0.411$, $p = 0.914$. As for valence users could provide their subjective value of ‘arousal’ from ‘-4 - calm’ to ‘4 - highly activated’. Also in this case, not all the users selected all the categories (‘-4’ = 31, ‘-3’ = 37, ‘-2’ = 36, ‘-1’ = 29, ‘0’ = 35, ‘1’ = 28, ‘2’ = 23, ‘3’ = 17, ‘4’ = 17). Since we did not find any significant relation between intimacy and arousal, we did not perform any post hoc tests. With the plot of Figure 7, we show that arousal cannot be considered dependent on the perception of intimacy.

6.2 Intimacy and Smartphone Usage: Data Preparation

We evaluated the collected intimacy on smartphone usage variables. Since we collected intimacy through ESM questionnaires appearing randomly but uniformly during the day, we computed smartphone usage features within a time frame in which the user answered to an ESM questionnaire.

6.2.1 Usage Sessions and Usage Windows

To create usage statistics for each user, we performed two main steps: (1) we aggregated the screen events and applications used in *usage sessions*, and (2) aggregated usage sessions in *usage windows*, defined by time thresholds (Figure 8). To each *window* (and therefore to all the contained *sessions*) we associated the intimacy value that the user submitted to the ESM questionnaire.

The screen events *ON* and *OFF* delimit what we define as a *usage session*. In this interval, we logged the name of the apps currently used with a sampling rate of 10 seconds (enabling us to capture micro usage, as shown by Ferreira *et al.* [33]). For each *usage session*, we noted its duration in seconds. Due to reasons bound to the implementation of the logger, we could guarantee this sampling rate only for the 20 participants from Switzerland (out of 42 study participants), thus, we only refer to these users for this analysis.

We define a *usage window* (Figure 8) as an aggregation of consecutive usage sessions that are not separated by a *timeout* longer than X minutes. One or many *usage sessions* compose a *usage window*. To study different usage window aggregations, we repeated the analysis with $timeout = \{2.5, 5, 7.5, 10, 15\}$ minutes.

The majority of usage sessions lasts less than 250 seconds (4.2 minutes). After this threshold, we have a very long tail of very few longer sessions, so we decided to consider only usage sessions with a maximum length of 600s (10 minutes) [23], [34], [35]. This procedure prevents the creation of outliers in further steps of our analysis.

6.2.2 Smartphone Usage Features

From the above-described base aggregation of *usage sessions* in *usage windows* we extracted three usage features:

1. *session_duration*, i.e., the duration of a single interaction with the smartphone;
2. *number_of_sessions*, i.e., the number of interactions within a given time window, and
3. *session_kind*, describing whether users were glancing at the smartphone screen or actively using one or more applications.

Since each user has her/his way to interact with the phone [23], the features are *unique to each user*.

To define the first feature, *session_duration*, we grouped the session duration in three bins: ‘short’, ‘medium’, ‘long’. We leveraged the distribution of all the session durations and found the breaks for the three classes employing a quantile method. At the end of the procedure, each *usage session* has a label corresponding to its *session_duration* bin.

For the *number_of_sessions* per window, we grouped the counts of usage sessions per usage windows in three bins: ‘few’, ‘several’, ‘many’. As for *session_duration*, we leveraged the distribution of the variable. In this case, each *usage window* has one of the *number_of_sessions* bins as a label.

For the *session_kind* feature, we considered the kind of interaction of users with their smartphones, using Banovic *et al.* [2] approach. We considered all the sessions from each user and divided these sessions into three categories: ‘lock screen only’ sessions, ‘launcher only’ sessions, and ‘applications and phone’ sessions. We then computed the distribution of session durations of the three categories. By finding the best separations between these three distributions with a maximum likelihood approach, we split the session duration in 3 bins, called ‘glance’ (sessions in which the users look at the phone lock screen),

‘review’ (sessions in which the users unlock the phone and mainly interact with the home screen), and ‘engage’ (in which the users unlock the phone and use one or several apps or perform/receive a phone call) [2].

6.2.3 Aggregating ESM Beeps to Phone Usage Data

Finally, we assigned the intimacy z-scores to the corresponding *usage sessions*. We first located the *usage window* in which the study participants gave a beep answer. We labeled the *usage session* of the located *usage window* in which the users provided their input as *DURING* (Figure 8) while we labeled the sessions in the same *usage window* executed before and after the ESM answer, as *BEFORE* and *AFTER* (Figure 8). Then we associated the intimacy Z-score with all *sessions* of the located *usage window*. We did not consider all the *usage sessions* belonging to *usage windows* that were not containing any *DURING usage session* (in practice all the *usage windows* that were not containing any ESM answer and, therefore, no *BEFORE* and *AFTER* sessions). We denoted these *usage sessions* as *OUT* as shown in Figure 8 and we discarded their respective *usage windows*. We also excluded from our analysis the *DURING* sessions, since we recorded them when the users were answering a new ESM questionnaire. In Table 1 we show the number of total *usage window* for all the considered values of a *timeout*. The decrease in the total number of usage sessions is due to the deletion of windows containing one *usage session* of more than 600 seconds (10 min), as described above. In Table 1, we present the window and session durations; for increasing *timeout* values the window duration increases. Moreover, in Table 1 we also show the percentage of intimacy labeled *AFTER* + *BEFORE* sessions (the ratio *between AFTER* and *BEFORE* always remains between 40% and 45%). We excluded from our analysis the windows generated with timeout values 10 and 15 (last two rows in Table 1) since it was unreasonable to associate a single intimacy value to windows having such a long average duration.

6.3 Analysis of Smartphone Usage Variables: Results

For this analysis, we used the same techniques applied before Linear Mixed Model (LMM) [32], and Bonferroni corrected post hoc tests. Here we performed tests for each value of *timeout* (as explained above it defines the start and end of *usage windows* with values: 2.5, 5, 7.5 minutes). We explain how all the three smartphone usage variables *session_duration*, *number_of_sessions*, and *session_kind* relate to intimacy significantly.

The *session_duration* relates significantly to intimacy. As presented in Table 2 (LMM column) the effect is significant for all the values of the *timeout*, except for '2.5' min. All the 20 users present all the cases for the *session_duration* variable for each *timeout*: i.e., there are 20 'short', 20 'medium', and 20 'long' sessions.

The post hoc tests (Figure 9 and Table 2) show that when users are in **lower intimacy** (higher z-score), they tend to have **shorter sessions than when in higher intimacy** (lower z-score). Thus, users interact for a shorter time with their smartphones when their perception of intimacy is lower and for a longer time when their intimacy is higher. For the categories 'short' and 'long' the LMM estimated means intimacy of opposite sign for all the *timeout* values. The difference is significant for '5', and '7.5' values of a *timeout*. For the 'medium' category, there is a change of sign with the increase of the *timeout* value. The bigger the *timeout* value is, the higher the number of sessions in the window is (Table 1). The 'medium' category contains a mix of session durations that are not clearly belonging to 'short' or 'long' categories. For values between '5' and '7.5' the mean *intimacy* of 'medium' changes sign, indicating a different balance of (short) sessions with low *intimacy*, and (long) sessions with high *intimacy*.

The *number_of_sessions* relates significantly to intimacy. In all the *timeout* cases the effect is significant (Table 3, LMM column), that means that *number_of_sessions* are related to the intimacy level. Not all the users had all the three bins for each minimum distance: '2.5' min = 6 few, 20 several, 20 many; '5' min = 12 few, 20 several, 19 many; '7.5' min = 14 few, 19 several, 19 many.

The post hoc tests (Figure 10, Table 3) show that users tend to have **a higher number of interactions when in low intimacy** (higher z-score) and **a lower number of interactions when in higher intimacy** (lower z-scores). Apart from the '2.5' value, all the other cases present a significant difference between the extreme categories 'few' and 'many', indicating that the intimacy plays a role in the number of consecutive usage session users perform. We observe a similar pattern for '2.5', '5', and '7.5' *timeouts* between the categories 'several' and 'many'.

Finally, also, the *session_kind* relates significantly to intimacy. The effect of *session_kind* is significant for all the *timeout* except for '2.5' (Table 4, LMM column). All the users present all the cases for this variable for each class: 20 glance, 20 review, and 20 engage.

With post hoc tests (Figure 11, Table 4) we reveal that users tend to **glance quickly at the phone when in lower intimacy** (high z-scores) while they tend to be **engaged when in higher intimacy** (lower z-

scores). In lower intimacy, users tend to perform phone activities that require a minimum time of interaction; while in higher intimacy they engage more with their devices and perform activities that require longer interactions. This fact indicates that the activity carried out on the phone correlates with changes in the users' intimacy perception. Also, in this case, there is a significant intimacy mean difference between the extreme categories 'glance' and 'engage' for '5' and '7.5' *timeouts*. The 'review' category is only significantly different from 'glance' when the mean intimacy value becomes negative (Figure 11, at '7.5').

7 Predictive Modeling of Intimacy

We proved that intimacy is a contextual variable predictive for smartphone usage in different intimacy conditions. In this part of the paper, we exploit the knowledge acquired by the previous data analysis to present a machine learning approach able to predict the perceived intimacy of users. Since the objective is to predict intimacy to make it available as further contextual information when users are using their smartphones, we do not base our intimacy prediction directly on smartphone usage variables. It would be convenient, but not practical. We cannot depend on smartphone usage if we want to predict automatically intimacy because we would need to wait that users perform their usage sessions to detect then their intimacy. Discovering intimacy at such a late stage would not be beneficial.

We proved that the perception of intimacy depends on context elements such as the location, the number and the kind of people around the user. Researchers are investigating techniques to define the number of people around a user, but as of today, there are no accurate systems able to provide a reliable estimation. Most of these techniques require precise requirements for an infrastructure (*i.e.*, instrumenting user's context) that at the moment we are not able to fulfill or are resource-demanding [19]–[22]. Identifying, who the people around an individual are, is even more difficult from a ubiquitous computation perspective. Therefore, even if we have the data for 'number of people' and 'kind of people' for each participant in our studies, we have decided not to use these in the machine learning approach, but instead rely only on users' location data.

We divided the procedure into 5 phases: (1) data preparation, (2) features extraction, (3) transformation of intimacy variable from Classes to Ranks, (4) generation of data sets for classification, (5) testing and validating the intimacy model via three experiments: (i) per subject model via 10 cross validation (10CV)

train-test, (ii) create a model per each user and test them against the data of the other users, and (iii) evaluation of accuracy of the prediction model in new user study.

7.1 Data Preparation

To exploit the location data, during our user study we collected smartphone connected mobile network's cell IDs. This value is the 'cheapest' location-related variable we can obtain on a user's smartphone. We recorded these cell IDs with a sampling rate of one minute. The cell ID variable is the aggregation of the cell ID of the connected tower, the Location Area Code (LAC the identifier of the geographical area of the country in which the tower is), the Mobile Network Code (MNC the identifier of the network provider in the country), and the Mobile Country Code (MCC). There are unique cell IDs across the two users samples of US and EU (all 42 users). To further define the users' locations, we grouped cell IDs in *neighborhoods*.

7.1.1 Creating Neighborhoods of Cell IDs

To cluster the cell IDs into meaningful neighborhoods, we employed a two-step process based on Fanourakis and Wac algorithm [36]. The first step determined if two consecutive cell ID measurements are physically near each other by analyzing the cell oscillations that often occur in cellular communications due to smartphone/cell power change, cell load balancing or some other operator-specific policies. We define a cell oscillation event between two cells, A and B, as any event where a smartphone is in a fixed location, and it connects to cell A then to cell B and eventually back to cell A in a relatively short time. However, due to the relatively low sampling rate of our logger data, we must make an assumption: the connection cannot 'oscillate' between two neighborhoods in the span of four minutes (*i.e.* the user does not oscillate between two significant neighborhoods in the span of four minutes). Thus, we guarantee that any cell oscillation event of time span less than four minutes occurs while the user was in the same significant neighborhood. We do not take into account any cell oscillation occurring over a longer time span, which could mean that the user left the neighborhood, was mobile and got back to the same neighborhood. Thus, we form a graph where the cells are the nodes, and two cells are connected if a cell oscillation event links them. The second step in determining the cell ID neighborhoods is to find the *cliques* of the cell ID graph that we created in the first step. So, we define each neighborhood of cell IDs

as each maximal clique of the graph where all possible pairs of cell IDs of them have a connection due to a cell oscillation event.

It is important to note that a cell ID can belong to multiple neighborhoods. For this reason, to assign a neighborhood to each cell ID in a sequence, we must choose the neighborhood that best matches the recent cell IDs in the sequence. To do this, we assigned weights to each cell ID in each neighborhood based on a factor related to the number of occurrences of that cell ID globally and specifically in that neighborhood. We also assigned weights to the cell IDs in the sequence based on how recently we recorded them. To assign a neighborhood to a cell ID in the sequence, we then compared each weighted neighborhood cell ID with the weighted sequence cell ID and derived the most likely neighborhood for that measurement. Based on these neighborhood assignments we calculated statistics such as the mean continuous time spent, minimum time spent, maximum time spent, and variance of the time expended in each neighborhood. Due to this procedure and technical problems on the retrieval of CellIDs from users smartphone, we could not include six users out of the 42 in the machine learning approach.

7.1.2 The Neighborhoods Cell IDs data

Each user has his list of neighborhoods, and each neighborhood has its unique ID, the mean, variance, min, max time spent in that neighborhood, its frequency, and its size. Each cell ID is part of a neighborhood. By mapping the cell ID to its neighborhood, we can know which was the neighborhood at that given time for each user. Therefore instead of cell ID per minute we assign each user his/her neighborhood ID per minute. Finally, we normalized all the statistics listed above across all the users by scaling them between 0 and 1 (we divided each variable by its own maximum). This operation allowed us to extract location related features from the statistics of neighbors to estimate the perception of the intimacy of each user.

7.2 Features Extraction

From the users' neighborhood data, we extracted the features for each user independently. We have assigned a neighborhood to each ESM answer. Once we identified the neighborhood we took the current scaled statistics of the neighborhood as features (mean, variance, max time spent in that neighborhood, its frequency, and its size). We also selected three more features from the statistics of the previous neighborhood (the neighborhood in which the user was before the current one), namely: mean, variance

of time spent in that neighborhood, and its frequency. To summarize, now for each user there are in total eight features for each ESM intimacy answer (left column of Table 5).

Since we envisioned to base the model for the intimacy prediction on a tree approach we discretized the features in bins. Some tree algorithms can deal with continuous variables, but an adequate discretization can help to obtain more accurate results. Our goal is to produce models that are representative of the possible different group of users in our dataset. Therefore, we derived the bins for each feature by considering all the users together. To create the bins we employed the unsupervised discretization filter (in Weka), automatically computing an adequate number of bins for each feature. Furthermore, since we have planned to use a tree-based approach and we know that most of the tree algorithms are not able to preserve the bin ordering [37] (p. 315). Therefore, to maintain the meaning of the order of bins, we set up the filter to transform each discretized attribute into a set of binary attributes. As described in [37] (p. 315) that we quote: “if the discretized attribute has k values, it is transformed into $k - 1$ binary attributes. If the original attribute’s value is i for a particular instance, the first $i - 1$ of these new attributes are set to false, and the remainder is set to true. In other words, the $(i - 1)$ th binary attribute represents whether the discretized attribute is less than i .” Therefore we transformed the initial 8 continuous features into 11 discrete features: `mean_time_bin{1,2,3}`, `var_time_bin{1,2}`, `max_time_bin{1,2}`, `frequency_bin{1,2}`, `size_bin{1,2,3}`, `prev_mean_time_bin{1,2,3}`, `prev_var_time_bin{1,2}`, and `prev_frequency_bin{1,2}`. In Table 5, we summarize the original continuous features and their binned counterparts. Each continuous variable corresponds to different intervals and number of bins for a final total of 19 features.

7.3 Transforming Intimacy Data: From Classes to Ranks

In [38] Martinez *et al.* showed that a ranking approach is more efficient when dealing with ordered and subjective data (i.e., on a scale, as intimacy from 1 to 6). Therefore, we focused our work on the Ranking by Pairwise Comparison (RPC) ([39] Hüllermeier *et al.*) classification approach. The RPC classification approach allows to classify ranks, and it fits our needs accordingly to [39]. These ranks can be preferences (e.g., A is preferred over B) or elements of an ordinal scale. The intimacy values are in fact part of an ordinal scale. This approach allows mapping users’ answers to an ordered preference scale, with at the first place the choice made by the user when answering the intimacy question. For RPC, the

ordered scales or preferences are expressed as $A > B$, which means that A is preferred over B for a given instance.

We created the following mapping between the single class value of intimacy to the complete ordered preference of the scale: '1' = '1>2>5>6', '2' = '2>1>5>6', '5' = '5>6>2>1', '6' = '6>5>2>1'. We omitted the ratings '3' and '4'. We did that for two reasons: (1) it is hard to define a clear ranking in the case of the ratings '3' and '4' (e.g., which is the rank preferred for '3'? '3>4>5>6>2>1' or '3>2>1>4>5>6', etc.) and (2) in [38] Martinez *et al.* suggested to remove the neutral values and '3' and '4' are representing those in our data. We then replaced the classes with these ranks in all the instances. We present here in Figure 12 and Figure 13, the distribution of the *intimacy* class (*i.e.*, ratings of ESM answers) and the ones for the derived rankings. The rankings represent the same distribution of the respective rating classes.

7.4 Generating Data Sets for RPC Classification

Martinez *et al.* in [38] also explain some techniques to reduce the subjective noise due to user errors on answering ESM survey questions or changes in users habits and attitude during the experiment. In particular, they propose to establish a window of instances that we can compare to each other inside the data of the same user. This approach allows to find cases in which the instances are similar (features values comparison), but finally, they present different or even opposite ranks. Therefore, we chose to create 6 different datasets: full, 'population_clean', 'user_clean_5', 'user_clean_10', 'user_clean_20', and 'user_clean_50'. The first dataset, denoted as full, is the full dataset as it is. For the second dataset, denoted as population_clean, we cleaned the data at population level considering all the user instances together. In this case, we analyzed the frequency of rankings represented by similar instances. We grouped the instances depending on the similarities of their features and kept counts of the frequency of the four rankings ('1>2>5>6', '2>1>5>6', '5>6>2>1', '6>5>2>1'). For each group of instances we summed up the counts of the two most similar ranks ('1>2>5>6' + '2>1>5>6' and '5>6>2>1' + '6>5>2>1'). We deleted from that group (and therefore from the full dataset) all the instances having the minor sum.

For the last four datasets, we cleaned instances at the user level with different window Y sizes for future instance look up, respectively $Y = \{5, 10, 20, 50\}$ (we denote these datasets as *user_clean_Y*). For these

datasets, we checked for noise in each user dataset independently using a moving window of Y instances. We cleaned out instances using the same principle adopted for the *population_clean* dataset we presented above.

In Table 6, we show the percentage of the removed instances for each resulting dataset and how their distribution per rank. In total, from 21474 instances from all datasets we removed 1539 (7.2%) of instances (removed distribution per rank: ‘1>2>5>6’ = 12.2%, ‘2>1>5>6’ = 27.2%, ‘5>6>2>1’ = 45.2%, ‘6>5>2>1’ = 15.4%). We removed a small part of the data, with the maximum excluded rate in the *population_clean* dataset (14.5%).

7.5 Testing and Validating the Intimacy Model: the Three Experiments

To test the potential of the RPC_based intimacy prediction model we performed three experiments: (1) per subject modeling, (2) subject model vs. all the other subjects instances, and (3) the model in practice with a user study in the wild. Each experiment contributes to the model refinement. We explain their goals, detailed configuration, results, and we discuss what is their contribution to the overall research objective. In each experiment, the accuracy of the model is measured by Kendall’s tau rank correlation ranging from -1 and 1 [39]. If the Kendall’s tau is closer to -1 the ranks predicted by the model are inverted with respect to the expected ones, instead if the Kendall’s tau is closer to 1 the rank outputted are the ones expected (high-quality result). A value of Kendall’s tau around 0 means a bad classification (no correlation between ground truth rankings and predicted ones).

7.5.1 Experiment 1: per Subject Modeling

The goals of this first experiment are: (a) evaluate the validity of our assumptions about the relation of time and place with the perception of intimacy for a significant number of users of our study; (b) understand the effect of the different datasets used for training on the accuracy of the prediction per each subject.

The configuration of the experiments is as follows. For each user independently and each dataset, we performed ten runs (ten different random seeds) of 10CV of RPC with ZeroR algorithm as an internal classifier (always predicting the majority class) with binary voting for the ranking. We use this setup to establish the baseline accuracy of our model for each user. We compare the first set up with ten runs (ten

different random seeds) of 10 cross validation (10CV) of the RPC algorithm using the J48 tree algorithm as an internal classifier with binary voting for the ranking.

The RPC J48 classifier performs significantly better than RPC ZeroR (baseline). The best dataset is the population_clean set. In Figure 14 we present the Kendall's tau means across all the users for the datasets and the two RPC classifiers, ZeroR, and J48. We performed two-way repeated measurements ANOVA to verify if the reported results are significant. We report: (1) the difference between the baseline classifier RPC ZeroR and RPC J48 is significant $F(1, 33) = 14.22$, $p = 0.001$, $r = 0.28$; (2) the dataset effect on accuracy of the model (mean Kendall's tau) is significant $F(1.4, 46.4) = 8.32$, $p = 0.003$, $r = 0.45$ (Greenhouse-Geisser correction); (3) The population_clean is the most predictive dataset, but is significantly better only over the full dataset ($p = 0.006$).

Focusing on the population_clean dataset, in Figure 15 we show the mean Kendall's tau of the ten 10CV runs of RPC ZeroR and RPC J48 for each user for the population_clean dataset. 25 users (76%, out of 33) using that datasets show an improvement over the baseline, five users (15%) have no improvement over the baseline, and three users have worse results (9%).

We can conclude that RPC J48 is significantly better than RPC ZeroR (baseline) and that the full dataset is the least suitable to generate a reliable intimacy model. Although the population_clean dataset is the best, we cannot exclude that the other cleaning method can lead to good results in practice.

7.5.2 Experiment 2: per Subject Modeling vs. Other Subject Data

The goals of the second experiment are: (a) identify clusters of users to investigate if it is necessary to create multiple models depending on some users particularities; (b) verify if different datasets generate substantial differences on the assumption (a).

The configuration of the experiments is as follows. For each user independently and each dataset, we based the intimacy model on his/her data and tested this model on the sets of data of all the other users (one by one). We constructed all the users models using experiment one RPC J48 with binary vote configuration.

In Figure 16 we present how all the users behaved against each single models of other users across all the datasets. We can observe the presence of two clusters: a majority of users have a high average performance (higher Kendall's tau mean), and some present a low accuracy (low Kendall's tau mean).

In Figure 17 we show the performance of each user but separated between datasets. We do this to highlight that the *population_clean* dataset is still the one achieving the highest accuracy, and all the other datasets are resulting in the same accuracies. We expected this outcome since the *population_clean* dataset generation required the harmonization of all the users' data (cleaning of outliers at the population level). However, that in any dataset, we can still note we have the presence of the two clusters of users mentioned above.

From Figure 18 we can notice that the *population_clean* dataset is performing significantly better than any other dataset. To confirm this fact we executed a one-way repeated measurements ANOVA and we report the following results: (1) the accuracy of the intimacy model is significantly affected by the dataset we used for the prediction $F(1.10, 37.35) = 25.98, p < 0.001, r = 0.06$ (Greenhouse-Geisser correction); (2) Bonferroni post hoc tests revealed that only the *population_clean* dataset is significantly better than all the other datasets ($p < 0.001$ for all).

To identify the clusters of users, for all the dataset independently we run a two-step clustering algorithm with automatic selection of the number of clusters. In Table 7 we present the clustering results. We have very similar results for all the datasets except for *population_clean* that being the best dataset presents different mean Kendall's tau centroids and less equilibrium on the number of users in the two clusters (~70% of users in the well-classified cluster, higher mean Kendall's tau). For each dataset, we have an average cluster silhouette of 0.7 (except *population_clean* 0.8), which means that the cluster quality is high, and the clusters of users are well separated. We can conclude that there exist two groups of users, which indicates that we need to create at least two different models for each dataset. The dataset *population_clean* is again the significant best, but it showed as well the presence of two groups of users, and some of them are affecting its results negatively as for the other datasets.

7.5.3 Experiment 3: Evaluating Model's Accuracy via User Study 'in the Wild'

The goals of the third experiment are: (a) to establish how accurate is the model in the real world (in the wild); (b) to observe if the models derived from the two clusters result in different accuracy; (c) verify that *population_clean* dataset is the most accurate intimacy model; (d) collect more ground truth to possibly refine further the intimacy model.

We configured the experiment to involve real smartphone users and run the algorithm on their own devices. First, we split all the datasets into the two clusters of users as output from the two-step clustering

procedure and created the respective models using Weka. We now have 12 models in total, two per each dataset. Second, we created an Android OS application implementing the neighborhood algorithm we presented in the previous sections and the prediction logic. Third, we created a series of notifications with which the users could provide their inputs and get informed about the predictions of the intimacy model. These notifications are like ESM, and in this context they are denoted as Ecological Momentary Assessment (EMA); they are providing the ground truth data.

In Figure 19 we show the EMA notifications users were receiving on their smartphones. For the first week of the study, we displayed to users the notification A (Figure 19). They were not required to do anything. We used the first days to collect data about the users' routine and generate the statistics for the neighborhoods at the base of our approach. After the first week, we started to predict the intimacy state of users. We performed a prediction every time we detected a change in the user context. Specifically, whenever there was a change of neighborhoods we were predicting the intimacy state with all the 12 models and creating a final prediction using a majority vote. At each prediction, we notified users with notification B (Figure 19). This notification allowed them to provide their intimacy input, from *very high* to *very low*, following the intimacy ranking of our model (1. *very high* > 2. *high* > 5. *low* > 6. *very low*). Once users provided their input, we were telling them the result of our majority vote prediction using notification C (Figure 19, two example of correct predictions). All the notifications provide to users a help button to instruct them about their meaning and tasks to perform. Finally, given the availability of new users' input, we added a module to collect extra data like user activity (*e.g.*, walking, biking, staying still), top WiFi access points, and light in lumen to investigate new possible variables that could help us to refine further the intimacy model. We collected users activity with the Google Services Android API. Each time we were notified by the service of a new activity we were logging it in a file. We collected WiFi access points every minute and the light from the smartphone sensor every time the user was switching on the smartphone screen (thus with a high probability the light value corresponds to the light when users provided their intimacy ground truth).

We involved a total of 31 users from our mQoL living lab in Geneva, Switzerland. Each of them contributed with one month of participation in the study. We collected a total of 3369 intimacy inputs and performed a total of 60307 predictions (5.6% have ground truth). In Figure 20 we present how each user contributed to the ground truth (EMA answers) and in Figure 21 we show the number of predictions we

performed per user. As for the previous user study, the number of contributions is not uniform. Some users are contributing more, some less. In some cases, the contribution is bounded by the small number of predictions. This little number of predictions can mean that the user was almost always in the same context, or our neighborhood algorithm was not able to detect changes.

The intimacy ranks from participants are distributed as follow: ‘1>2>5>6’ = 1249 (37%), ‘2>1>5>6’ = 770 (23%), ‘5>6>2>1’ = 628 (18.6%), ‘6>5>2>1’ = 722 (21.4%). As the intimacy ranks from our predictions are different for each rank and model, we provide them in the analysis, as follows.

Using the Kendall’s tau ranking correlation metric we compared each users’ input to the corresponding 12 models outputs and the respective majority-voting outcome. We present the results of this analysis in Figure 22 with a heat map in which we display the mean Kendall’s tau across all the users, for each model and the outcome of the majority voting. For each model, we have the two variants we created with the data originating in the users’ cluster 1 (C1) and cluster 2 (C2), as we presented in experiment 2. For some users, we are successfully predicting the intimacy state as in the theoretical model (mean Kendall’s tau from 0.4 to 0.6), while for some other users (e.g., users 7, 19, and 26) appositve rankings are observed (mean Kendall’s tau from -0.4 to -0.6), and some are around 0. For users with appositve rankings, the explanation is that the classifier inverted their intimacy rankings. Theoretically, a solution could be to invert these rankings to obtain a better accuracy, but in practice, it is hard to detect these inverted ranks after the classification and react accordingly. The most concerning users are the ones having a mean Kendall’s tau around zero (e.g., users 1, 4, and 25). For these, our model accuracy is low; there is no correlation at all between their ranking and the one our model predicts.

We have investigated the raw data and results and we observe that the main reason behind the Kendall’s tau around zero is that our intimacy models predicted the rank ‘6>5>1>2’ (e.g., majority-voting 2607 predictions out of 3369 marked with ground truth) instead of the rank ‘6>5>2>1’ (recalling the expected distribution of 722 (21.4%)). The models were also unable to predict the rank ‘1>2>5>6’ (e.g., majority-voting 0 predictions out of 3369, expected distribution was 1249 (37%)). These two problems are at the origin of a low-rank correlation of users that were mostly in rank ‘1>2>5>6’. The users that obtained better results are the ones that were primarily in rank ‘2>1>5>6’ (distribution of 770 (23%)). For this rank, we have a distribution for the majority-voting of 761 predictions. A deeper analysis is needed to

identify the reasons behind these misclassifications. We comment further on these results in the Discussion section.

As we show in the heat map in Figure 22, there are not significant variations of each model Kendall's tau within each user. To evaluate the different models globally, in Figure 23 we provide the mean Kendall's tau of each of the models across all the users. We did not find a significant mean Kendall's tau difference between the models. We executed a one-way repeated measurements ANOVA: (1) the model variable is not significantly affecting the prediction results $F(2.10, 63.05) = 1.302, p = 0.280$ (Greenhouse-Geisser correction); (2) Bonferroni post hoc tests revealed that only the *subject_50_c1* model is significantly better than *population_c1* ($p = 0.035$).

Among the additional context data that we collected on users' smartphone, we found that user activity, WiFi, and light correlate with intimacy. For each of the three variables, we mapped each user intimacy vote to the closest available state of them. For example for activity, starting from the user vote timestamp we searched for the most recent previous activity state. To analyze activity and WiFi data we performed a count of the possible states of the two variables and the current intimacy state. For activity, in Figure 24 we show a heat map with the percentages of the counts normalized across each user and the four intimacy ranks, for four activities groups: 'vehicle/biking', 'still', 'other' (unknown + tilting), and 'on foot'. We can notice that for very high intimacy ($1 > 2 > 5 > 6$) all the users (apart 1, 2, 18, and 22) tend to be mostly in the 'still' state. The lower the intimacy, the more mobile are the users. In very low intimacy ($6 > 5 > 2 > 1$) for the majority of the users we have more activities indicating mobility. We performed a Cochran-Mantel-Haenszel Chi-Squared Test for count data confirmed that these differences are significant: $M^2 = 227.91, df = 9, p < 0.001$ (for this test we removed 8 users: 1, 2, 8, 9, 16, 18, 22, and 27, because they did not have sufficient votes for each intimacy category).

For WiFi, we created four categories as follows. "No_wifi" represents all the cases in which users were not connected to wifi, "first_wifi" is the wifi to which the users were connected most of the time, "second_wifi" is the second most used WiFi, and "other" is the collection of the remaining WiFi. In Figure 25, we present a heat map showing the percentage of each wifi category normalized per each user and each intimacy state. The main pattern highlighted in the graph is the high presence of "no_wifi" when users are in a low intimacy ($5 > 6 > 2 > 1$ and $6 > 5 > 2 > 1$). Vice-versa "first_wifi" is particularly high in the very high intimacy state ($1 > 2 > 5 > 6$). Being connected to the most used WiFi can mean high intimacy and

not be connected to WiFi can mean low intimacy. We performed a Cochran-Mantel-Haenszel Chi-Squared Test for count data confirmed that these differences are significant: $M^2 = 363.88$, $df = 9$, $p < 0.001$ (as before, for this test we removed 8 users: 1, 2, 8, 9, 16, 18, 22, and 27).

Finally, for the ‘light’ context variable, we analyzed how the mean normalized light value (lumen) changes in the four intimacy states. For each user, we transformed their absolute light values with z-scores (we set their mean lumen to 0). In Figure 26 we show a heat map with the four intimacy states, and the mean scaled lumen for each user. From the graph, we can see a possible relation between being in low intimacy and having higher mean lumen values. This fact can mean that when users are in high intimacy, they are probably inside where lumen values are lower than outside (when users are more probable to be in low intimacy). A One-Way Repeated ANOVA confirmed this assumption: (1) depending on the intimacy rank we have significantly different mean values of lumen $F(1.79, 39.46) = 4.101$, $p = 0.028$ (Greenhouse-Geisser correction); (2) Bonferroni post hoc tests revealed that when users are in intimacy rank ‘1>2>5>6’ their smartphone light sensors captured a lower lumen level than when in intimacy ranks ‘2>1>5>6’ ($p = 0.024$) and ‘5>6>2>1’ ($p < 0.001$). We found no differences between the other ranks. Also in this case, we excluded all the users not having light values for all the four intimacy states.

8 Discussion

8.1 Validation of Intimacy Concept

In the results we show that the three objective context elements: ‘place’, ‘number of people’, and ‘kind of people’ around correlate with users’ intimacy. In particular, for the location, we have shown that users perceive ‘home’ differently than any other less familiar place. For the number of people, as the number of people increases, the perception of *intimacy* decreases. Finally, for the kind of people, we identified that being with the closest ones (*e.g.*, family and friends) leads to higher *intimacy* than when with non-closest ones (*e.g.*, strangers and co-workers). The user study results show that our definition of *intimacy* is close to what it can represent to the users in reality (considering the three objective context elements: location, number and kind of people around).

8.2 Intimacy and Valence

We explored whether the subjective context mood elements like valence and arousal relate with intimacy. Despite the low significance of the results (due to the high number of valence categories and the total number of answers from users), valence seems to be related to intimacy, while arousal does not appear to have any particular relation. The relationship with valence suggests that intimacy can enclose the users' feelings regarding the perceived comfort of the situation. However, valence is only one of the many subjective variables, and it is particularly applied to mood and stress metrics. To achieve more accurate results, more subjective context elements should be analyzed and compared to intimacy.

8.3 Quick Smartphone Glances Means Low intimacy and Engaging

Sessions Means High Intimacy

From the literature [6], [8], [9], we have hints that *the intimacy perception can influence the behavior of people*. Our results indicate that users have different interaction patterns with their smartphone when their perception of intimacy changes. Users perform shorter, more interrupted (high frequency of sessions), and less engaging tasks in lower intimacy. Vice versa, they perform longer, more continuous (less frequent sessions), and more engaging tasks when in higher intimacy.

These changes in behavior correlate with the subjective perception of the environment (in our case described by three objective contextual elements). When the perception of intimacy is high, users are usually in a place like home, with fewer and usually only close people. This environment facilitates the users (1) to feel more secure [6], and perform certain smartphone interactions that they would not perform elsewhere (that may require more attention, precision, concentration, and trust); (2) to be less distracted [8], and complete in “a single shot” the intended smartphone interaction and tasks; (3) to be willing to share their experience directly and indirectly with other people present [9], and engage more with their devices.

8.4 The Subjective Perception of Intimacy Can be Predicted and Operationalized in Practice

The results of our machine learning approach show that by leveraging features extracted from smartphone and related only to the spatial/temporal context, specifically related to locations visited by users, it is

possible to predict the user’s intimacy obtaining significantly better results than the baseline. This accuracy has the potential to be increased further by integrating features like the number of people around the user and possibly the kind of people. To achieve that, it would not be necessary to predict the exact number or the exact kind of people around the user, but it will be required to identify some features able to describe the situation the user is in as a whole. We are aware that several works were performed to study how to detect people around [19]–[22], and it would be certainly useful to derive from them a combination of variables able to describe this phenomenon. Moreover, if we also consider the mood component, valence, we may be able to integrate further new features. Additionally, we identified a correlation between intimacy and activity, WiFi, and light that we need to investigate and probably integrate into the prediction model. We hypothesize that adding these features to the location ones will further increase the accuracy of our prediction method.

The practical deployment of the model ‘in the wild’ showed that it is feasible to implement what we envision theoretically, but many variables need to be refined to make the system fully operational. Particularly, in practice, there is probably no need to create multiple models since no one proved to be significantly more robust than others in practice. Most probably the most accurate solution is to re-generate a single model by repeating the same procedure presented in this paper with the new data acquired in the latter user study. Another aspect to verify is why the models were not predicting the higher intimacy rank accurately ($1 > 2 > 5 > 6$). Our first assumption is that the underlining algorithm for the neighborhood detection was failing to recognize some places or we had challenges in the computation of statistics for the different places, the users visited. This last assumption requires further research on the neighborhood algorithm.

Our second assumption is that when we generated the different models we biased them towards predicting low levels of intimacy. The solution can be to balance the instances used to train the future models and produce a more balanced predictor.

8.5 Limitations

The ESM/EMA method itself is the main limitation on capturing the users’ intimacy. With this approach, we may have lost some users’ rare events or situations (Lathia *et al.* [31]), in particular when the users happened to be in low intimacy states, as our data indicates. Thus, some particular situations may be

missing. For subsequent studies, location-based features could be used to predict when to trigger ESM/EMA questions, since, as we have shown in our research, the place is an accurate descriptor of *intimacy* and its changes are easily identifiable.

Another limitation stems from the choice of options we proposed in the different ESM questions (*e.g.*, for number of people: ‘alone’, ‘2-10’, ‘40+’). It is hard to design these in advance, especially because, to our knowledge, the literature does not clearly present validated scales suitable for the purpose of our study.

Based on this work, in subsequent works, we may refine the possible ESM questions and answers.

Moreover, the six points Likert-scale we used may not be the best representative for the *intimacy* states.

The data distribution may suggest the adoption of visual analog rating scales with high and low categories at the extreme. Finally, due to the large number of possible combinations, we did not verify how the factors ‘place’, ‘number of people’, and ‘kind of people’ interact with each other to then influence *intimacy*.

The limitations of the intimacy modeling relate mostly to the availability of a high number of good instances to generate useful predictive models. It is hard to collect data on the specific situations and the intimacy ground truth without specific users study. Additionally, the subjectivity of the intimacy perception makes the prediction procedure very user specific. It is difficult to create a model that work well at the population level.

9 Conclusions and Future Work

Our future work focuses on two main directions: (1) further understanding on how we could leverage intimacy in the design of mobile user interface and interaction design choices and (2) implementing a complete system able to estimate the current user’s perception of intimacy and evaluate the impact of intimacy as a subjective context element variable in interface design in a real world application.

9.1 Intimacy and Interface Design

Given the potential for the *intimacy* discussed in this paper, we can define the following design implications for pervasive systems and for each of them we present an example of how high and low intimacy levels may influence them:

9.1.1 Notifications

Service providers and applications developers can notify the users (*e.g.*, via a notification on the smartphone) about promotions, particular events or another kind of notifications. For example, when in *high intimacy* users are more likely to spend more time on their device once we capture their attention. In this case a marketing campaign, which requires some user actions (*e.g.*, answer a survey to win a prize or post a Twitter with a specific hashtag to get a discount coupon), would be probably more successful because the users will be more likely to engage. If an application issues this kind of notification when the user is in *low intimacy*, it would be probably forgotten or the task not completed (at least if the task does not allow a very fast or intermittent interaction).

9.1.2 Shortcuts

Particular applications can adapt their interface to offer easy tasks or shortcuts when users are in *low intimacy*. For example, in *low intimacy* a newsreader application may propose to the user fewer more targeted and shorter articles. Instead in *high intimacy* it may provide some possible matches (*e.g.*, based on the “if you like that, you may also like this” approach) that requires users to be more open to exploration and therefore engage more with the application. Another example can be a bus trip planner applications that in *low intimacy* (most probably the user is in the street) proposes only the very first and immediate departures or the ones that the user takes usually. Instead, in *high intimacy*, it would show more departures because most probably the user would engage more with the app and explore other alternatives.

9.1.3 Reduce Intrusiveness

Some services may reduce their intrusiveness depending on the users’ intimacy. For example, in case we would prove the relation of intimacy with privacy requirements, when users are in *high intimacy* an application or service may automatically avoid the collection of some personal context information (*i.e.*, in *high intimacy* context elements may be more familiar and valuable to the users). This fact would require a further study on how users value other context information when they are in *low* and respectively in *high* intimacy.

9.1.4 OS improvements

The smartphone itself may adapt its functionalities depending on the users' intimacy. For example, in *high intimacy* to never lock the device or change the way notifications are delivered (e.g., depending on users preferences on how they would like to be "interrupted" in *high* or *low intimacy* situations). Another example, in *high intimacy* prefetching of data, may be suspended at OS level (to optimize battery level), because users would probably be more relaxed and do not require high responsiveness than when in *low intimacy* when usually they need to have faster responses and actions to their requests.

9.1.5 Further User Behavior Exploration

Application designers may use *intimacy* to analyze different unexplored users' behaviors. In this paper, we discussed the smartphone usage and the intimacy of users relation in a high-level fashion. It is highly possible, as indicated already in the examples we just presented above, that research groups design and execute more experiments in particular fields to verify the pertinence of intimacy with particular situations. For example, as we stated before, we may need to check the relevance of the connection between privacy and intimacy to automate the non-disclosure of particular contextual information in certain intimacy conditions. Another important task can be to see if there exist real benefits to modify the user interface of applications based on the intimacy of users. Are the users really in need of such optimizations?

On our side, to investigate the effectiveness of these possible improvements based on intimacy, we will conduct larger user case studies. This approach may require the implementation of a real smartphone application where we are going to test some of the above use cases. One significant challenge of such approach is to test the effectiveness of intimacy unobtrusively. We will have to collect the current user's perception of intimacy to perform the required dynamic changes to the user interface (and other interaction elements). This system will be effective only when our intimacy prediction model will reach a satisfactory level of accuracy and efficiency regarding battery consumption. However, we can already sketch and envision a possible experiment in which we can test intimacy in practice.

9.2 Intimacy in Practice

The focus of our future work is to refine the intimacy model with a cycle of prototypes and experiments. The final goal is to integrate the intimacy metric in a mobile public transport application offering real

time bus departures (being developed by the lab, currently operational) [40]. The study will consist of ‘A/B’ tests where half of the user will experience app interface changes when their intimacy state changes and the other half will not. We will compare the results gathered in the two groups to identify the effects of the intervention. For example, it will be done by estimating the user experience with the app derived from the app engagement data such as average usage time sessions and length of task execution. We may define other measures to verify the accuracy of the *intimacy* model while users perform usual tasks in the transport application. For example, we may introduce a new feature consisting in the identification of the 2 to 3 most intimate user’s places. We will identify the intimate places with our intimacy model. We will then ask the users to label these places with *home*, *work* or a personalized place or to delete them if they are not relevant. This procedure can enable to verify the accuracy of our algorithm in identify intimate places and, therefore, being able to assess correctly the users’ high intimacy for those situations. We may then use these places to provide to users several statistics about their movements between these places (e.g., the average duration of the trip depending on the crowd, time of the day and line used).

Acknowledgements

We would like to thank Professor Anind Dey for his help, feedback, and the opportunity to conduct a user study in his HCII Ubicomp lab at Carnegie Mellon University (Pittsburgh, PA, USA) and Professor Marc Langheinrich (Ubiquitous Computing group at the Università della Svizzera Italiana) for his valuable inputs.

Our research has been supported by the Swiss National Science Foundation Project "Enabling People-Centric Sensing by Overcoming the privacy BarriEr" (PCS-OBEY) [SNSF-149591], European Ambient Assisted Living project "A Pervasive Guardian for Elderly with Mild Cognitive Impairments" (AAL-MyGuardian) [AAL-2011-4-027], and European Ambient Assisted Living project "intergenerAtioNal communitY for coMpAny knowledge TransfEr" (AAL-ANIMATE) [AAL-2013-6-071].

References

- [1] V. Woollaston, “How often do YOU look at your phone? The average user now picks up their device more than 1,500 times a week,” *MailOnline*. [Online]. Available: <http://www.dailymail.co.uk/sciencetech/article-2783677/How-YOU-look-phone-The-average-user-picks-device-1-500-times-day.html>. [Accessed: 03-Feb-2015].

- [2] N. Banovic, C. Brant, J. Mankoff, and A. K. Dey, "ProactiveTasks: the Short of Mobile Device Use Sessions," in *MobileHCI*, 2014.
- [3] K. Prager, *The psychology of intimacy*. The Guilford Press, 1997.
- [4] O. Dictionary, "Intimacy definition," 2014. [Online]. Available: <http://www.oxforddictionaries.com/definition/english/intimacy>.
- [5] O. Dictionary, "Intimate definition," 2014. [Online]. Available: <http://www.oxforddictionaries.com/definition/english/intimate>.
- [6] L. C. Manzo, "Beyond house and haven: toward a revisioning of emotional relationships with places," *J. Environ. Psychol.*, vol. 23, no. 1, pp. 47–61, Mar. 2003.
- [7] D. Seamon and J. Sowers, "Place and Placelessness, Edward Relph," *Key Texts Hum. Geogr.*, no. Seamon 2000, pp. 43–51, 2008.
- [8] S. Saegert, "Crowding: Cognitive Overload and Behavioral Constraint," *Environ. Des. Res.*, pp. 254–260, 1973.
- [9] E. Diener and M. E. P. Seligman, "Very Happy People," *Psychol. Sci.*, vol. 13, no. 1, pp. 81–84, Jan. 2002.
- [10] R. S. Miller and H. M. Lefcourt, "The assessment of social intimacy.," *J. Pers. Assess.*, vol. 46, no. 5, pp. 514–8, Oct. 1982.
- [11] I. Altman, *The Environment and Social Behavior: Privacy, Personal Space, Territory, and Crowding*. ERIC, 1975.
- [12] A. F. Westin, *Privacy and Freedom*. Bodley Head, 1970.
- [13] R. S. Gerstein, "Intimacy and privacy," *Ethics*, vol. 89, no. 1, pp. 76–81, 1978.
- [14] G. Chen and D. Kotz, "A survey of context-aware mobile computing research," *Tech. Rep.*, pp. 1–16, 2000.
- [15] C. Bolchini, C. Curino, and E. Quintarelli, "A data-oriented survey of context models," *ACM Sigmod*, vol. 36, no. 4, pp. 19–26, 2007.
- [16] C. Bettini, O. Brdiczka, K. Henriksen, J. Indulska, D. Nicklas, A. Ranganathan, and D. Riboni, "A survey of context modelling and reasoning techniques," *Pervasive Mob. Comput.*, vol. 6, no. 2, pp. 161–180, Apr. 2010.
- [17] J. Ye, S. Dobson, and S. McKeever, "Situation identification techniques in pervasive computing: A review," *Pervasive Mob. Comput.*, vol. 8, no. 1, pp. 36–66, Feb. 2012.
- [18] A. K. Dey, "Understanding and using context," *Pers. Ubiquitous Comput.*, vol. 5, no. 1, pp. 4–7, Feb. 2001.
- [19] N. Eagle and A. (Sandy) Pentland, "Reality mining: sensing complex social systems," *Pers. Ubiquitous Comput.*, vol. 10, no. 4, pp. 255–268, Nov. 2005.
- [20] N. Eagle, A. S. Pentland, and D. Lazer, "Inferring friendship network structure by using mobile phone data.," *Natl. Acad. Sci. USA*, vol. 106, no. 36, pp. 15274–8, Sep. 2009.

- [21] B. Adams, D. Phung, and S. Venkatesh, "Sensing and using social context," *ACM Trans. Multimed. Comput. Commun. Appl.*, vol. 5, no. 2, pp. 1–27, Nov. 2008.
- [22] J. Zheng and L. M. Ni, "An unsupervised learning approach to social circles detection in ego bluetooth proximity network," *Proc. 2013 ACM Int. Jt. Conf. Pervasive ubiquitous Comput. - UbiComp '13*, p. 721, 2013.
- [23] H. Falaki, R. Mahajan, S. Kandula, D. LyMBERopoulos, R. Govindan, and D. Estrin, "Diversity in smartphone usage," in *MobiSYS*, 2010, p. 179.
- [24] T. Soikkeli, J. Karikoski, and H. Hammainen, "Diversity and end user context in smartphone usage sessions," *Int. Conf. Next Gener. Mob. Appl. Serv. Technol.*, pp. 7–12, 2011.
- [25] R. Likamwa, Y. Liu, N. D. Lane, and L. Zhong, "MoodScope: Building a Mood Sensor from Smartphone Usage Patterns," in *MobiSYS*, 2013, pp. 389–401.
- [26] G. Bauer and P. Lukowicz, "Can smartphones detect stress-related changes in the behaviour of individuals?," in *Percom*, 2012, no. March, pp. 423–426.
- [27] M. Gustarini and K. Wac, "Smartphone Interactions Change for Different Intimacy Contexts," in *Mobicase*, 2013, pp. 1–18.
- [28] M. Bradley and P. Lang, "Measuring emotion: the self-assessment manikin and the semantic differential," *Behav. Ther. Psychiat*, vol. 25, no. 1, pp. 49–59, 1994.
- [29] J. M. Hektner, J. A. Schmidt, and M. Csikszentmihalyi, *Experience Sampling Method: Measuring the Quality of Everyday Life*. SAGE Publications, Inc, 2007.
- [30] R. Larson and P. Delespaul, "Analyzing Experience Sampling data: a guidebook for the perplexed," *Exp. Psychopathol.*, 1992.
- [31] N. Lathia, K. K. Rachuri, C. Mascolo, P. J. Rentfrow, and U. Kingdom, "Contextual Dissonance : Design Bias in Sensor-Based Experience Sampling Methods," in *UbiComp*, 2013, pp. 183–192.
- [32] G. Verbeke and G. Molenberghs, *Linear Mixed Models for Longitudinal Data*. Springer, 2009.
- [33] D. Ferreira, J. Gonçalves, V. Kostakos, L. Barkhuus, and A. K. Dey, "Contextual Experience Sampling of Mobile Application Micro-Usage," in *MobileHCI*, 2014.
- [34] A. Oulasvirta, T. Rattenbury, L. Ma, and E. Raita, "Habits make smartphone use more pervasive," *Pers. Ubiquitous Comput.*, vol. 16, no. 1, pp. 105–114, Jun. 2011.
- [35] T. Yan, D. Chu, D. Ganesan, A. Kansal, and J. Liu, "Fast app launching for mobile devices using predictive user context," in *MobiSYS*, 2012, p. 113.
- [36] M. Fanourakis and K. Wac, "Lightweight Clustering of Cell IDs into Meaningful Neighbourhoods," in *HET-NETs*, 2013, pp. 698–704.
- [37] Ian H. Witten, Eibe Frank, and M. A. Hall, *Data Mining Practical Machine Learning Tools and Techniques*. 2011.
- [38] H. Martinez, G. Yannakakis, and J. Hallam, "Don't Classify Ratings of Affect; Rank them!," *Affect. Comput.*, vol. 5, no. 3, pp. 1–14, 2014.

- [39] E. Hüllermeier, J. Fürnkranz, W. Cheng, and K. Brinker, “Label ranking by learning pairwise preferences,” *Artif. Intell.*, vol. 172, no. 16–17, pp. 1897–1916, Nov. 2008.
- [40] M. Gustarini, J. Marchanoff, M. Fanourakis, C. Tsiourti, and K. Wac, “UnCrowdTPG : Assuring the Experience of Public Transportation Users,” in *The 2nd IEEE WiMob 2014 International Workshop on Smart City and Ubiquitous Computing Applications*, 2014, pp. 1–7.

Tables

	Windows				Sessions			
timeout	Total	# avg sessions	Avg	SD	Total	Avg	SD	Intimacy labeled
(min)	number	per window	duration	duration	number	duration	duration	<i>AFTER + BEFORE</i>
			(min)	(min)		(min)	(min)	(%)
2.5	18748	25.5	12.4	10	33327	0.89	1.41	7
5	13484	37.6	26.3	19.3	31669	0.85	1.40	13
7.5	10716	48.8	43.8	33.3	29982	0.80	1.32	20
10	8786	61.4	73.3	59.7	28449	0.76	1.28	27
15	6555	84.9	151.7	135.2	25284	0.74	1.26	36

Table 1: Extracted phone usage data statistics about usage windows and usage sessions.

timeout (min)	LMM results	short vs. medium	short vs. long	medium vs. long
2.5	$F(2, 43.808) = 2.17, p = 0.126$	$0.04, p = 1$	$0.18, p = 0.138$	$0.14, p = 0.481$
5	$F(2, 40.581) = 5.26, p = \mathbf{0.009}$	$0.09, p = 0.363$	$0.23^a, p = \mathbf{0.011}$	$0.14, p = 0.244$
7.5	$F(2, 42.094) = 7.79, p = \mathbf{0.001}$	$0.11, p = 0.940$	$0.23^a, p = \mathbf{0.002}$	$0.18, p = 0.242$

^a. The mean difference is significant at the 0.05 level with Bonferroni correction included

Table 2: LMM analysis and pairwise comparisons results for *session_duration* for each *timeout*.

timeout (min)	LMM results	few vs. several	few vs. many	several vs. many
2.5	$F(2, 9.110) = 15.06, p = \mathbf{0.001}$	$0.14, p = 1$	$-0.41, p = 0.442$	$-0.55^b, p < \mathbf{0.001}$
5	$F(2, 23.137) = 9.27, p = \mathbf{0.002}$	$-0.25, p = 0.238$	$-0.46^b, p = \mathbf{0.008}$	$-0.21^b, p = \mathbf{0.036}$
7.5	$F(2, 28.738) = 14.3, p < \mathbf{0.001}$	$-0.25, p = 0.106$	$-0.50^b, p = \mathbf{0.001}$	$-0.25^b, p = \mathbf{0.001}$

^{a.}

The mean difference is significant at the 0.05 level with Bonferroni correction included

Table 3: LMM analysis and pairwise comparisons results for *number_of_sessions* for each *timeout*.

timeout (min)	LMM results	glance vs. review	glance vs. engage	review vs. engage
2.5	$F(2, 29.166) = 3.19, p = 0.056$	-0.28, $p = 1$	0.16, $p = 0.076$	0.19, $p = 0.290$
5	$F(2, 31.300) = 8.45, \mathbf{p = 0.001}$	0.09, $p = 0.560$	0.21 ^c , $\mathbf{p = 0.001}$	0.12, $p = 0.276$
7.5	$F(2, 32.900) = 8.84, \mathbf{p = 0.001}$	0.15 ^c , $\mathbf{p = 0.045}$	0.20 ^c , $\mathbf{p = 0.001}$	0.05, $p = 1$

^{a.}

The mean difference is significant at the 0.05 level with Bonferroni correction included

Table 4: LMM analysis and pairwise comparisons results for *session_kind* for each *timeout*.

Original Continuous Feature	Binary Binned Features
mean_time	mean_time_bin1 $\{(-\infty - 0.27] \text{ or } (0.27 - \infty)\}$
	mean_time_bin2 $\{(-\infty - 0.733] \text{ or } (0.733 - \infty)\}$
	mean_time_bin3 $\{(-\infty - 0.976] \text{ or } (0.976 - \infty)\}$
var_time	var_time_bin1 $\{(-\infty - 0.2585] \text{ or } (0.2585 - \infty)\}$
	var_time_bin2 $\{(-\infty - 0.9375] \text{ or } (0.9375 - \infty)\}$
max_time	max_time_bin1 $\{(-\infty - 0.4205] \text{ or } (0.4205 - \infty)\}$
	max_time_bin2 $\{(-\infty - 0.9535] \text{ or } (0.9535 - \infty)\}$
frequency	frequency_bin1 $\{(-\infty - 0.23] \text{ or } (0.23 - \infty)\}$
	frequency_bin2 $\{(-\infty - 0.969] \text{ or } (0.969 - \infty)\}$
size	size_bin1 $\{(-\infty - 0.2795] \text{ or } (0.2795 - \infty)\}$
	size_bin2 $\{(-\infty - 0.801] \text{ or } (0.801 - \infty)\}$
	size_bin3 $\{(-\infty - 0.9615] \text{ or } (0.9615 - \infty)\}$
prev_mean_time	prev_mean_time_bin1 $\{(-\infty - 0.2465] \text{ or } (0.2465 - \infty)\}$
	prev_mean_time_bin2 $\{(-\infty - 0.709] \text{ or } (0.709 - \infty)\}$
	prev_mean_time_bin3 $\{(-\infty - 0.976] \text{ or } (0.976 - \infty)\}$
prev_var_time	prev_var_time_bin1 $\{(-\infty - 0.155] \text{ or } (0.155 - \infty)\}$
	prev_var_time_bin2 $\{(-\infty - 0.9375] \text{ or } (0.9375 - \infty)\}$
prev_frequency	prev_frequency_bin1 $\{(-\infty - 0.1725] \text{ or } (0.1725 - \infty)\}$
	prev_frequency_bin2 $\{(-\infty - 0.969] \text{ or } (0.969 - \infty)\}$

Table 5: Original continuous features and their corresponding binary binned features.

Dataset	Kept	Removed	Removed distribution for ranks			
			'1>2>5>6'	'2>1>5>6'	'5>6>2>1'	'6>5>2>1'
full	3579 (100%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
population_clean	3059 (85.5%)	520 (14.5%)	7 (1.3%)	61 (11.7%)	299 (57.5%)	153 (29.5%)
user_clean_5	3376 (94.3%)	203 (5.7%)	46 (22.7%)	78 (38.4%)	67 (33%)	12 (5.9%)
user_clean_10	3348 (93.5%)	231 (6.5%)	47 (20.3%)	78 (33.8%)	89 (38.5%)	17 (7.4%)
user_clean_20	3309 (92.5%)	270 (7.5%)	45 (16.6%)	93 (34.5%)	109 (40.4%)	23 (8.5%)
user_clean_50	3264 (91.2%)	315 (8.8%)	42 (13.4%)	108 (34.3%)	132 (41.9%)	33 (10.4%)

Table 6: The datasets generated from the six different cleaning techniques with their percentages of kept and removed instances.

Dataset	# Clusters	Clusters Distribution (Users per Cluster)	Clusters Centroids (Mean Kendall's Tau)	Cluster Quality (Average Silhouette)
full	2	18 (51.4%) / 17 (48.6%)	0.38 / 0.12	0.7
population_clean	2	24 (68.6%) / 11 (31.4%)	0.53 / 0.19	0.8
subject_clean_5	2	19 (54.3%) / 16 (45.7%)	0.38 / 0.10	0.7
subject_clean_10	2	18 (51.4%) / 17 (48.6%)	0.39 / 0.10	0.7
subject_clean_20	2	21 (60%) / 14 (40%)	0.37 / 0.07	0.7
subject_clean_50	2	21 (60%) / 14 (40%)	0.37 / 0.07	0.7

Table 7: The separation of users in clusters for each of the six datasets with the mean Kendall's Tau of each cluster centroid and the average Silhouette value for cluster quality (> 0.5 represents a good cluster separation).

Figures

1234567?

A

AIDE / HELP

VERY HIGHHIGHLOWVERY LOW?

B

AIDE / HELP

PREDICTIONYOUR INPUT?

Very LowVery Low

PREDICTIONYOUR INPUT?

HighHigh

C

AIDE / HELP

62

Captions

Figure 1: Number of users answers to ESM questions. Some users are contributing more than others creating a heterogeneous number of contributions.

Figure 2: The different intimacy scores mean per each study participant (1 ‘completely’ intimate - 6 ‘not at all’ intimate) [error bars 95% CI].

Figure 3: Mean intimacy z-score for places. Lower is the value of the mean z-score, higher is the user intimacy [error bars 95% CI].

Figure 4: Mean intimacy z-score for number of people. Lower is the value of the mean z-score, higher is the user intimacy [error bars 95% CI].

Figure 5: Mean intimacy z-score for kind of people. Lower is the value of the mean z-score, higher is the user intimacy [error bars 95% CI].

Figure 6: Mean intimacy z-score for valence. Lower is the value of the mean z-score, higher is the user intimacy [error bars 95% CI].

Figure 7: Mean intimacy z-score for arousal. Lower is the value of the mean z-score, higher is the user intimacy [error bars 95% CI].

Figure 8: An example of usage window containing the *DURING* session (including the user answer to the ESM - beep), the *BEFORE*, and *AFTER* sessions. The usage sessions labeled as *OUT* are in usage windows without ESM answer.

Figure 9: Estimated mean z-score of context intimacy for the short, medium, and long *session_duration* for each *timeout* [error bars 95% CI].

Figure 10: Estimated mean z-score of context intimacy for the few, several, and many *number_of_sessions* for each *timeout* [error bars 95% CI].

Figure 11: Estimated mean z-score of context intimacy for the glance, review, and engage *session_kind* for each *timeout* [error bars 95% CI].

Figure 12: Distribution of the intimacy states over all the users.

Figure 13: Distribution of intimacy ranks after transformation from intimacy scores.

Figure 14: Mean Kendall’s Tau across all the users for the datasets and the two RPC classifiers, ZeroR (baseline) and J48 for the 10CV test of experiment 1 [error bars 95% CI].

Figure 15: Mean Kendall's tau of the ten 10CV runs of RPC ZeroR (baseline) and RPC J48 for each user for the population_clean dataset [error bars 95% CI].

Figure 16: The mean Kendall's Tau for each user tested against the models of other users over all the datasets [error bars 95% CI].

Figure 17: The mean Kendall's Tau for each user tested against the models of other users for each dataset. The dashed black is the population_clean dataset that performs better than the others [error bars 95% CI].

Figure 18: The mean Kendall's Tau for all the datasets over all the run of users against other users model. The dataset population_clean is significantly better than the others [error bars 95% CI].

Figure 19: The notification we presented to the users during experiment 3. Notification A is the countdown of the seven days of data acquisition. With notification B users could provide their current intimacy state. We were displaying B each time the intimacy model was computing a new prediction. Notification C was presenting the users input vs. our prediction.

Figure 20: The number of users answers to the notifications of experiment 3. As for the first user study, some users contributed more than others.

Figure 21: The number of predictions that our model did for each user involved in experiment 3. Some had a higher number of predictions probably due their higher mobility.

Figure 22: Mean Kendall's Tau heat map for all the 12 datasets (c1 datasets are generated by users cluster one and c2 with users cluster two) and majority vote (from the 12 datasets outcome). A positive mean Kendall's Tau means we correctly predicted the intimacy ranks; a zero means there is no correlation (wrong prediction), and a negative one represents an inverse ranking prediction.

Figure 23: Mean Kendall's Tau over all the users for all the datasets. No datasets seem to be significantly better than the others [error bars 95% CI].

Figure 24: Heat map showing the counts in percent of the users activities in the different intimacy rankings. The counts are normalized per rank and user. Users tend to be less mobile in high intimacy than in low intimacy.

Figure 25: Heat map showing the counts in percent of the users WiFi in the different intimacy rankings. The counts are normalized per rank and user. Users tend to be connected to a frequent wifi when in high intimacy and not connected to WiFi when in low intimacy.

Figure 26: Heat map showing the mean light (in Lumen) for all the users in the different intimacy rankings. The light is normalized with z-scores for each user. Light seems to be overall more intense when users are in low intimacy (probably outdoor) than when they are in high intimacy.