Please cite this article as: L. Kqiku and D. Reinhardt, Lengthy Early Morning Instant Messages Reveal More Than You Think: Analysing Interpersonal Relationships using Mobile Communication Metadata, Pervasive and Mobile Computing (2023) 101781, https://doi.org/10.1016/j.pmcj.2023.101781.

Lengthy Early Morning Instant Messages Reveal More Than You Think: Analysing Interpersonal Relationships using Mobile Communication Metadata *

Lindrit Kqiku[1,a], Delphine Reinhardt[a]

[a] *Computer Security and Privacy, University of Göttingen, Göttingen, Germany*

Abstract

Privacy policies are one of the key factors that determine user's privacy exposure. Quite often such policies request users' consent to "only" collect their metadata. In this paper, we investigate the information that communication metadata can reveal, specifically, in regards to users' social circles. To this end, we developed an application and conducted a longitudinal field study with 25 participants, who installed our application on their personal smartphones. Over a period of four weeks, our application collected the metadata of the participants' communication with their social contacts over four channels, i.e., calls, SMS, e-mails, and *Instant Messages* (IMs). The content of the communication were not collected and the identities of the participants' social contacts were encrypted, so that only the users could get access to them. We leverage the collected metadata to examine whether and to what extend it is possible to exploit them in order to classify the participants' social contacts into four social categories, namely, family members, friends, acquaintances, and colleagues using *Machine Learning* (ML) tech-

*The final publication is available at Elsevier via `https://doi.org/10.1016/j.pmcj.2023.101781`

[1]Goldschmidtstr. 7, 37077 Göttingen, Germany, Phone: +49 551 39-172063, Fax: +49 551 39-14403, E-Mail: kqiku@cs.uni-goettingen.de

niques. By doing so, we do not only reproduce and replicate an existing study, but also extend it by further considering the metadata about IMs. In our user study, friends and family members call each other and exchange SMS more than acquaintances and colleagues. Moreover, as expected, IMs are exchanged more with friends followed by family, whereas e-mails more between acquaintances and colleagues. In addition, to validate the role of instant messaging channel, we show against our expectations that considering metadata about IMs only slightly improve the prediction of users' social ties. We also examine the most important features that lead to such predictions. We show that the prediction of social ties can be further enhanced by considering the aforementioned communication channels. Our study results in the f-measure score of 79.4% for a fine-grained classification of four considered social categories, achieving better results than related approaches, and 89.6% when considering the family category against all the others.
Keywords: Interpersonal Relationships, Privacy, Android, Metadata

1. Introduction

Since January 2021, WhatsApp users are pushed to accept a new privacy policy [1]. The main update of the policy is the requirement to accept sharing the user's metadata with Facebook. Sharing these metadata goes beyond the already existing collection of metadata about users' activity, such as their phone unique identifier and location information by WhatsApp [1]. In a recent study [2], the authors show that approximately half of the surveyed users are unaware about the detailed meaning of the term metadata itself. Such data, which might be perceived as innocuous by users at the first glance, may indeed reveal sensitive information about them, such as their communication patterns and their interpersonal relationships. This

information can in turn reveal additional sensitive information about their health [3, 4] or well-being [5, 6], for example.

Among potential inferences based on communication metadata, we especially focus on interpersonal relationships. We build upon existing studies [7, 8] and extend them according to two dimensions. We first replicate and confirm the results we obtained in [8] and further consider metadata about outgoing IMs, a dimension that has been ignored until now.

Our contributions can be summarized as follows:

1. We designed and implemented an Android application that log users' contact list and metadata about calls, SMS, e-mails, and incoming/outgoing IMs. To collect IMs metadata, we introduced a specific logger for WhatsApp, Facebook Messenger, Telegram, and Threema. Our choice is motivated by their respective market shares in Europe [9, 10].
2. We further designed and conducted a user study, in which 25 participants installed our Android application after having been informed about the study modalities and left it installed for four weeks in average. As a result, we collected metadata about calls, SMS, and e-mails from all of them. For 15 of them, we also collected their metadata about their communication using different instant messengers. The participants further labeled their contacts according to the following social categories: colleagues, friends , family, and acquaintances.
3. We thoroughly analyze the resulting dataset by considering a set of features derived from the aforementioned channels, by (1) excluding IMs to verify the results obtained in [8] and (2) including features derived from instant messaging metadata to investigate their added value. In a second step, we apply different classifiers and consider

different numbers of social categories, in which the participants' social contacts are classified to compare our results with [7, 8, 11]. We further apply feature selection techniques to determine the most important features that help the most to classify the participants' social contacts in the correct category.

4. We finally present the obtained results and compare them to those obtained in [8] (without IMs) and [11] (with 8 participants and 249 labelled contacts vs. 25 resp. 15 participants with 2,499 resp. 1,484 contacts in our case).

Our results show that e-mails are exchanged more between acquaintances and colleagues rather than family members and friends. In contrast, IMs are exchanged more with friends and family members. Overall, our approach leads to a higher f-measure score for fine-grained classifications as opposed to the related approaches by 20.9% [8] (with five categories), respectively by 9% (with four categories) [11]. We hence show that a set of four pre-defined social relationship categories lead to good results as opposed to five of them as in [8]. We also show that Random Forest model is the best fitted model for our dataset and the one in [8] provided by the authors, although not considered by them. Our results further demonstrate that user's social ties could be classified with a precision score of 80.5% for the four considered categories, resp. 86.3% for the three categories, and 92.3%, with regard to two. We further demonstrate that our models are only slightly improved, with 1.4% increase in f-measure score, when IMs metadata are included.

Our gained insights can be beneficial for different domains. For example, being able to classify users in different categories can be leveraged in access control schemes, particularly for online sharing decisions. As pro-

Table 1: Comparison of our approach with existing works based on considered (meta)data. Prox: Proximity, Loc: Location, Pic: Pictures, Demo: Demographics

References	Calls	SMS	E-mails	IMs	Prox	Loc	Pic	Demo
[19]						x	x	
[20]					x			
[21]	x	x				x		x
[22]				x	x			
[23]	x	x				x		
[7, 24–27]	x	x						
[8]	x	x	x					
[11]	x	x		x		x		
Our approach	x	x	x	x				

posed in [12], supporting users in choosing the appropriate audience by making suggestions based on their own data could address the shortcomings of existing solutions demonstrated in multiple studies [13–18].

Our paper is structured as follows. We discuss related work in Sec. 2. We detail our application in Sec. 3, our dataset in Sec. 4, and the classification in Sec. 5. We discuss our results in Sec. 6 and conclude in Sec. 7.

2. Related Work

Different works [7, 8, 11, 19–27] aim at inferring social ties using different types of data and metadata. Tab. 1 provides an overview of the different data types leveraged for this purpose and compare our approach with others. In more details, our approach is especially grounded on the following prior works that consider fine-grained social ties. In the work of Min et al. [7], the authors designed and implemented models to predict social ties according to three categories, namely, *family, coworkers,* and *social contacts.* Whereas, in our previous work [8], we further divided the latter category, ending up with

friends, acquaintances, uni/school mates, co-workers, and *family.* The work sharing the most similarities with this paper is [11], in which metadata of calls, SMS, and IMs are used to determine the users' relationship according to *friend, family, work, and hobby* categories. Note that they used a multi-label user labelling process, i.e, a contact could be labelled with more than one categories. They however considered only WhatsApp and Threema as instant messaging channels. The IMs were further collected from status bar notification and they were limited to only incoming messages. Moreover, they did only consider a subset of the eight participants' contacts, leading to only 249 instances. As opposed to our work, they hence did not consider e-mails, outgoing IMs, and all labelled contacts in their modellings.

In summary, in contrast to the aforementioned studies, we particularly explore incoming and outgoing instant messaging metadata from WhatsApp, Facebook Messenger, Telegram, and Threema, besides calls, SMS, and e-mails. We further consider two to four social categories.

3. Logging Application

We designed and implemented an Android application to collect metadata from calls, SMS, e-mails, and IMs. Tab. 2 shows these metadata. Note that caller and callee in the case of calls, as well as, sender and receiver, in the case of other data channels, are encrypted. We also ever store any content data. For metadata about calls, SMS, we used Android native content providers. We adopted JavaMail API and e-mail content providers for e-mails. For IMs, we applied the following approach.

3.1. Collecting IM Metadata

We decided to collect metadata from the three most represented messengers, namely WhatsApp, Facebook Messenger, and Telegram, considering their widely use [9, 10]. We also included Threema for its privacy-friendliness to cover participants who may be more privacy-aware. We selected it over other privacy-friendly applications because of its higher market share in Europe [28]. In all selected applications, the data were collected using the Android accessibility service as detailed in what follows.

3.1.1. Android Accessibility Service

This service is originally intended for people with special needs. Google provides several services based on this service, e.g., screen reader for blind users. The accessibility service API is also open to developers for new applications. Instead of using it according to its original purpose, we used it to collect the IM metadata, which could not be collected otherwise. In a nutshell, the accessibility service works in the background and receives callback information from the system when accessibility events are triggered. Examples of such events include clicking on a list view or any other event that triggered using the *User Interface* (UI). We used the main classes of

Table 2: Collected metadata. Receiver type, in the case of e-mails, includes "to", "cc", and "bcc", whereas it consists of types such as "text", "location", "replies" for IMs.

	Caller or sender	Callee or receiver	Receiver type	Duration or length	Type	Date	Emoji counter
Calls	x	x		x		x	
SMS	x	x		x			
E-mails	x	x	x	x	x	x	
IMs	x	x	x	x	x	x	x

accessibility service including *accessibility events* and *accessibility node info*. With the former, we listen to events that occur in the background when an application is being used. We use the latter to access the information visible in the active window of the UI. As an output, *accessibility node info* returns the results from the window's content in a tree structure.

To distinguish the relevant metadata from other information, we analysed how the selected messengers operate. This means that we analysed the logging data that the accessibility service accesses when the respective messenger is used in all scenarios, such as going from unknown to conversation directly and vice versa, or going in settings tab after conversation mode. We also analysed all possible types of information of each messenger, e.g., text, images, images with description, and reply. We then analysed and implemented different methods to distinguish these cases. All relevant information that are triggered by the accessibility service are captured by a method specific to the underlying messenger. The gathered information are next included in a *Finite State Machine* (FSM) (as a finite automata), which switches between finite states when triggered by a particular event.

We next briefly detail the specific design and implementation developed for each messenger to log the corresponding metadata. We further highlight the encountered challenges and how we solve them.

WhatsApp. We designed a WhatsApp FSM with four main states:

$$S_{i=4} \epsilon \{unknown, home, to\_conversation, conversation\}$$

The FSM is triggered when Whatsapp is first used. Its state changes from *unknown* to *home* when a main activity event is triggered and detected in the background by the accessibility service.

Figure 1: A simplified overview of the WhatsApp messages logger and the corresponding instant message table. # refers to numbers.

As depicted in Fig. 1, the chat list is displayed to the user in the *home* state (see Ⓐ). Once the user taps to chat, the state switches from *home* to *to_conversation* (see Ⓑ) followed by *conversation* (see Ⓒ).

In the *to_conversation* state, the latest interaction date and the conversation partner are extracted using an *accessibility event* method (see Ⓓ). The metadata are extracted from the content on the fly as they come or as they are shown on the active window conversation mode of the messenger. These metadata are then stored as a message object (see Ⓔ and Tab. 2) with the conversation partner and the date (see Ⓕ).

In the *conversation* state, the nature of the messages are extracted. In this state, the *accessibility event* catches the active window content and by calling *accessibility node info*, we obtain a *RelativeLayout* tree with *ViewGroups* consisting of various lengths. Different types of messages (e.g., text, voice, audio, reply, image, video, etc.) have different tree structures. Simi-

larly, both sent and received messages result in different tree structures. As a result, both the message types and direction can be easily distinguished.

For the message dates, particular care is required since its format depends on the chosen language. Therefore, we dynamically covered all cases. Besides, we converted incomplete date (e.g. "today"), or dates without years (e.g. *MMM dd* in US English) to a full date (e.g. *MMM dd, YYYY*).

Threema. The FSM and the metadata extraction is similar. The states, last interaction dates, and conversation partners are identified using *accessibility event* methods. Specific functions cover date formats and message types.

Facebook Messenger. In contrast to WhatsApp, extracting metadata of messages is more complex, but relies on a similar state machine.

Unclear state distinction. A clear distinction between the *home* to *conversation* state, i.e., the *to_conversation* available in WhatsApp, is not possible, as there is no event that define these states. We hence used *accessibility node info* to define them, instead of directly using the events derived from *accessibility event* methods. To this end, we looked at the structure of the *accessibility node info* for both states and differentiated between them based on their skeleton tree structure. Moreover, we used a similar method based on *accessibility node info* to find the part of the tree structure that contained the conversation partner and the date.

Date extraction issues. Additional adaptations were required due to multiple date and time formats. For example, if a message is sent within the last six days, it is written as a string of three letters, e.g, *FRI*. We hence implemented a function that compares today's date, transforms it to this string format, and compares the grabbed date against all possible days. This also applies for months and the current year. We hence covered all

possible format deviations to extract the correct date. Moreover, not all sent or received messages are directly stamped with the corresponding time. However, messages that are sent within the same day and within a short time period are grouped together and their time is stamped above them in the middle of the rendered view. We hence implemented a function to grab this time and stamp each message within this group.

Dynamic structures. Depending on how the message was rendered in the active window of the UI, the text child of *accessibility node info* tree could have a different structure: It could be positioned in one or two different locations. As a result, we covered this dynamic for all message types.

Telegram. We further adapted our solution to the specifics of Telegram.

In summary, we designed and implemented a specific solution for each messenger to collect the associated metadata. Our implementation by nature is dependent on the messenger version, which could have changed during our study. Using our app, users can further access the collected metadata and detailed statistics as well as stop the data collection.

4. Dataset Collection

We deployed our app in a field study in January and February 2021.

4.1. Study Design and Settings

The study was approved by our data protection officer. We made sure to minimize potential harm to our participants. Their participation was voluntary and they could opt out at any time. We have informed the participants by providing an extensive description of the collected data.

Once they had agreed, the participants were assisted in installing and using our app on their own phone in a virtual session. For participants who

felt confident enough to configure and install the app, we also provided video and paper instructions in three different languages. The participants set a username and a password, so that they can later access their decrypted data. They were next presented a list of permissions to collect the different data types, which they could either accept or dismiss. If they accepted to collect e-mail metadata, they configured the access to their e-mail account(s).

The metadata were collected for about four weeks. They were periodically uploaded to a server hosted by our institution. The participants could change the upload frequency and stop the data collection in the app.

After the data collection period, the participants logged in the application again to access the list of people they had interacted with. We asked them to edit this list by (1) merging different identifiers for the same person and indicating irrelevant contacts as well as (2) labelling the remaining contacts. For the labelling, we asked the participants to classify each contact according to one of the given categories: *Acquaintances* (ACQ), *family* (FAM), *colleagues* (COL), and *friends* (FRI). Moreover, as in [8], an *ignore* category could be used if the contact could not be assigned to one category. The decision for these categories is grounded in the state-of-the-art (see Sec. 2). During this step, we remotely assisted the participants if necessary.

4.2. Participants

25 participants contributed to this study. They all own an Android phone and use at least one considered messenger. We recruited them within our and their social networks. We took care of recruiting a diverse sample from different demographic profiles. Their age ranges between 18 and 65 years. Out of 25, 16 are males and 9 are females. They are all based in European countries incl. Kosovo, Germany, and Switzerland.

Table 3: Number of contacts per each social category and distribution of the collected metadata per communication channel and social category for both S_{25} and $S_{15(IM)}$.

Class	#	Calls	SMS	E-mails	Σ
ACQ	549	2,299	1,791	1,939	6,029
COL	1054	3,441	1,078	1,321	5,840
FAM	383	13,701	7,880	1,113	22,694
FRI	513	6,499	5,676	540	12,715
Σ	2,499	25,940	16,425	4,913	47,278

(a) S_{25} dataset

Class	#	Calls	SMS	E-mails	IMs	Σ
ACQ	684	1,289	434	1,691	635	4,049
COL	266	1,452	463	780	1,131	3,826
FAM	249	6,937	4,369	413	2,501	14,220
FRI	285	4,787	3,934	212	3,349	12,282
Σ	1,484	14,465	9,200	3,096	7,616	34,377

(b) $S_{15(IM)}$ dataset

4.3. Resulting Datasets

We obtained two datasets presented in Tab. 3. The first dataset S_{25} includes the metadata obtained for calls, SMS, and e-mails of all 25 participants (see Tab. 3a). In contrast, $S_{15(IM)}$ includes metadata about IMs in addition to the other communication channels for 15 participants (see Tab. 3b). We observe that the participants in S_{25} communicated most with their family members followed by friends using calls and SMS. In contrast, more e-mails have been exchanged with acquaintances and colleagues than the other social categories. The same trends are observable in $S_{15(IM)}$. Beside, IMs are predominantly used with friends and family. Tab. 4 presents the extrema and quartiles of the number of logged events over the different channels. Calls are the most preferred communication channel.

Table 4: Minimum, quartiles, mean, and maximum for each communication channel and participant-wise in $S_{15(IM)}$ and S_{25}

	Calls		SMS		E-mails		IM
	$S_{15(IM)}$	S_{25}	$S_{15(IM)}$	S_{25}	$S_{15(IM)}$	S_{25}	$S_{15(IM)}$
Min	6	6	5	1	1	1	4
Q_1	214	160	78.5	26	14	25	165
Q_2	962	735	132	127	34	122	664
Mean	964.3	1,038	613.3	657	266.3	393.5	585.8
Q_3	1457	1,434	293.5	317	366.9	366	808
Max	2,751	5,203	2,957	4,314	1,200	1,529	1,385.3

5. Classification

We next investigate whether and to what extent it is possible to identify the social ties based on the collected metadata.

5.1. Feature Extraction

Like to [7, 8], we consider the following factors to infer social ties: (1) intensity, (2) regularity, (3) temporal tendency, and (4) maintenance cost. The extracted features and time periods are presented in Tab. 5 resp. 6. They include 56 features for calls, 55 for e-mails resp. SMS, and 72 for IMs.

5.2. Resampling Methods

The number of contacts for each category is slightly unbalanced. To boost our results, we applied techniques to under-sample resp. over-sample the datasets. We herein focus on the best results obtained with the *Synthetic Minority Over-sampling Technique* (SMOTE) [29] of WEKA with five neighboring instances. We resampled the original dataset with SMOTE using uniform distribution according to the highest number of contacts per class. Note that to allow comparability with [11], we also applied resampling with replacement.

Table 5: Extracted features adapted from [7, 8]. #: number, DUR: duration, AVG: average, STD: standard deviation, lengthly-calls: duration > than 2x the average duration.

Logged data	Factors	Features
Calls	Intensity	Total {#, DUR}, total #lengthy-calls
	Regularity	{AVG, STD, MIN, MAX} # calls per week {last month, last half year, whole interval}, # days called/days logged, {AVER, MAX} DUR {all, outgoing, incoming}
	Temporal tendency	# and DUR weekend, weekday/total # or total DUR, {#, DUR} for each of the week/total # or DUR, {# and DUR} {early morning, morning, afternoon, evening, early night, late night}/total # or DUR
	Maintenance cost	{#, DUR} calls for past {2 weeks, 3 months} / total calls
SMS, e-mails	Intensity	Total {length, #} of messages
	Regularity	{AVG, STD, MIN, MAX} # {last month, last half year, whole period}, # days communicated/days logged, {AVG, MAX} length {all, outgoing, incoming}
	Temporal tendency	# and length weekend, workday/total # or total length, {#, length} for each of the week/total #, length, {# and length} {early morning, morning, afternoon, evening, early night, late night}/total #
	Maintenance cost	{#, length} for past {2 weeks, 3 months}/total {#, length}
IMs	Intensity	Total {length, #} of messages, Total # of emojis
	Regularity	{AVG, STD, MIN, MAX} # {last month, last half year, whole period}, # days communicated/days logged, {AVG, MAX} length {all, outgoing, incoming}
	Temporal tendency	# and length weekend, workday/total # or total length, {#, length} for each of the week/total #, length, {# and length} {early morning, morning, afternoon, evening, early night, late night}/total # of messages and emojis , # of weekend, workday emojis, # of emojis for each of the weekday
	Maintenance cost	{#, length} for past {2 weeks, 3 months}/total {#, length} of messages, and # for past {2 weeks, 3 months}/total # of emojis

Table 6: Matching between defined time periods and corresponding hours

Early morning	5:00 - 8:59	Evening	17:00 - 20:59
Morning	9:00 - 12:59	Early night	21:00 - 00:59
Afternoon	13:00 - 16:59	Late night	1:00 - 4:59

5.3. Categories

We further consider the following classification cases: (a) Four categories: ACQ, FAM, COL, and FRI, (b) Three categories: (ACQ $\cup$ FRI), FAM, and COL, and (c) Two categories: FAM vs. (ACQ $\cup$ FRI $\cup$ COL). The resulting

datasets are referred to as S_{x-y-z} with $x \in \{25, 15(IM), 15, [8], [11]\}$ being the reference of the original dataset (S_{15} corresponds to $S_{15(IM)}$ without considering IMs., y the resampling methods, i.e., SMOTE or resampling with replacement noted *replacement*, and $z \in \{2, 3, 4\}$ the aforementioned number of categories.

5.4. Evaluation and Results

We present our results for these categories and compare them to [8, 11]. For a comprehensive evaluation, we selected the following classification models: Random Forest, decision trees, SVM, and naive Bayes. After resampling, we performed 10x10-fold cross validation using WEKA. Note that we obtained similar results using the hold-out and resample with replacement techniques. We performed hyper-parameter tuning for potential performance improvement. The best performing algorithm was Random Forest. The number of trees was the main parameter that increased the performance. We obtained the best results with 200 iterations.

5.4.1. Classification based on Four Classes: ACQ, FAM, COL, FRI

Tab. 7 shows that the inclusion of IMs increases the performance of almost all models, except naive Bayes. For SVM, the precision decreases, but simultaneously the recall rate increases, thus resulting in an increased f-measure. This effect could be reduced by adjusting the confidence interval. Due to the overall lower performance of SVM, we did not further optimize it. The best results are obtained with Random Forest. By incorporating all communication channels, the f-measure is 79.4% with 1,484 labelled contacts. Removing a communication channel leads to a reduction of this measure in all cases, the most important reduction being observed when

Table 7: Comparison of classification performance for $S_{15(IM)-SMOTE-4}$

.

Evaluated channel	Original contacts	Algorithm	Precision (%)	Recall (%)	F-measure (%)	ROC (%)
IMs	1,484	Random Forest	80.5	78.5	79.4	93.7
		J48 Decision tree	62.1	63.7	62.8	78.2
		Naive Bayes	44.0	21.2	28.4	68.6
		SVM	34.1	71.1	40.3	58.0
No IMs	1,376	Random Forest	79.3	77.1	78.1	93.3
		J48 Decision tree	61.2	62.8	61.8	77.9
		Naive Bayes	36.7	40.5	38.3	69.2
		SVM	44.8	44.1	32.1	56.6
No SMS	1,386	Random Forest	78.4	77.0	77.6	93.2
		J48 Decision tree	62.0	61.7	61.7	77.6
		Naive Bayes	47.7	18.8	26.7	67.4
		SVM	40.8	59.4	38.7	58.2
No emails	861	Random Forest	64.8	70.0	67.1	87.3
		J48 Decision tree	47.9	49.9	48.6	67.9
		Naive Bayes	33.2	16.8	21.9	65.0
		SVM	27.8	84.8	41.8	55.6
No calls	1,171	Random Forest	80.0	75.3	77.5	93.2
		J48 Decision tree	62.6	62.6	62.4	77.8
		Naive Bayes	47.5	25.7	33.0	72.2
		SVM	39.1	70.7	42.2	60.9

Table 8: Performance of Random Forest using $S_{15(IM)-SMOTE-4}$ (one run).

Class	Precision (%)	Recall (%)	F-measure (%)	ROC (%)	COLL	ACQ	FAM	FRIENDS
COLL	84.8	77.7	81.1	94.3	531	86	32	34
ACQ	70.6	75.3	72.9	90.9	57	515	60	52
FAM	80.4	85.5	82.9	94.8	14	50	585	35
FRIENDS	81.4	77.6	79.5	91.6	24	78	51	530
Mean	79.3	79.0	79.1	92.9	Accuracy = 79.042%			

removing e-mails. However, note that removing a communication channel reduces the number of original contacts (e.g., 861 out of 1,484 for e-mails).

Tab. 8 shows that the classification rate is rather well balanced across the different classes ranging between 79.65% for ACQ and 77.78% for FAM.

We next compare the results obtained with $S_{25-SMOTE-4}$ as well as $S_{15(IM)-SMOTE-4}$ against $S_{[8]-SMOTE-4}$. Note that we collected the same metadata about calls, SMS, and e-mails in [8] of 19 participants in 2014. For comparison purposes and based on our previous results, we merged our previously distinct categories, *coworkers* and *work/schoolmates* in [8], into COL.

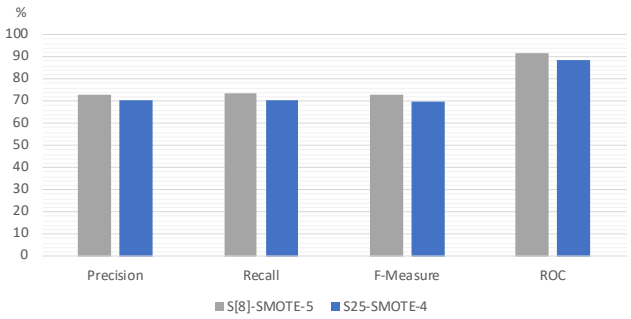

Figure 2: Comparison of the Random Forest model using f-measure metric.

We use Random Forest as basis for the comparison depicted in Fig. 2. Overall, the performance is better with $S_{[8]-SMOTE-4}$. This difference may be attributed to different participants' demographics or the evolution in terms of communication channels between 2014 and 2021. However, if we consider $S_{15(IM)-SMOTE-4}$, the f-measure score is higher than with $S_{[8]-SMOTE-4}$.

We next compare the results obtained for our dataset $S_{15(IM)}$ against the one in [11] that includes the metadata of calls, SMS, cell tower location, and the incoming WhatsApp and Threema IMs from eight participants. We apply in this case the same WEKA settings as in [11] (i.e., 70% for training and 30% for testing with resampling with replacement, over 10 runs by randomizing the dataset over each run). Their original WORK and HOBBY categories are equivalent to our COL and ACQ categories, respectively. Tab. 9 shows that higher f-measures are obtained with our dataset than the results presented in [11] for Random Forest. The highest differences are for FRI followed by COL and FAM. These differences could be attributed to the (1) inclusion of the e-mail metadata channel (i.e, better for classifying especially acquaintance category), (2) logging of outgoing IMs aside from incoming messages, (3) logging of more messengers, the (4) higher number of extracted features, and/or (5) a larger number of participants.

Table 9: Random Forest results for $S_{15(IM)-Replacement-4}$ and $S_{[11]-Replacement-4}$.

$S_{[11]-Replacement-4}$ [11]		$S_{15(IM)-Replacement-4}$	
Class	f-measure (%)	Class	f-measure (%)
WORK	63.8	COL	77.3
HOBBY	71.9	ACQ	70.7
FAM	68.5	FAM	77.8
FRI	60.9	FRI	75.2
Mean	66.3	Mean	75.3

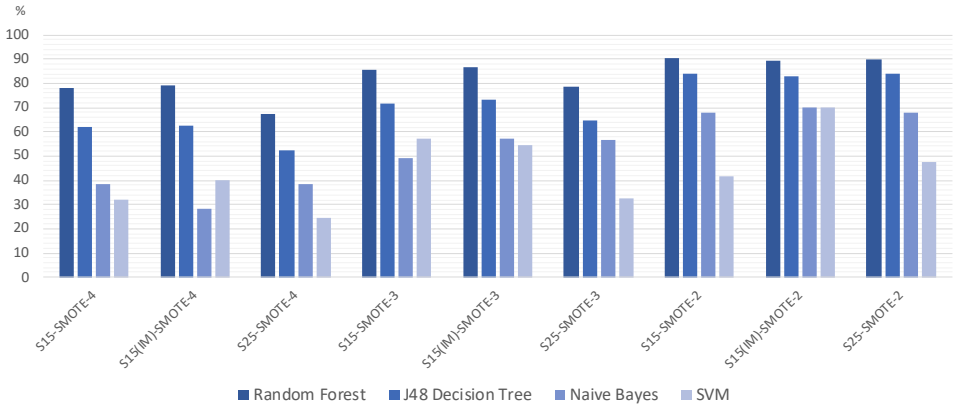

Figure 3: Comparison of different models for 2-4 social categories based on f-measure.

5.4.2. Classification based on three classes: (ACQ ∪ FRI), FAM, and COL

Like in [7, 8], we merged ACQ and FRI into one category. The corresponding results for three categories shown in Fig. 3 are better than those obtained with four categories. Again, Random Forest shows the best performance. Note that the best results were achieved with SVM in [8], though. The inclusion of IMs only improves the performance of Random Forest, Decision Tree, and Naive Bayes. This improvement remains limited, especially in the case of Random Forest and Decision Tree. This slight increase may be caused by two factors. (1) The participants communicate with their contacts over multiple channels, which already contributes to a correct clas-

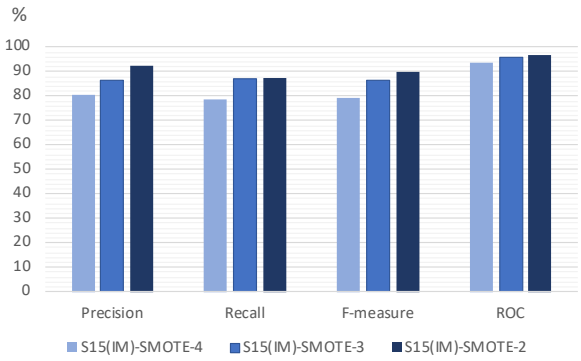

Figure 4: Random Forest performance for 2-4 social categories and all channels.

sification, the IMs just confirming them (i.e., out of 247 contacts identified using IM channel, 139 of them communicate through IM and at least another channel). (2) The participants communicate only over IMs with some contacts (i.e., out of 247 contacts identified using IM channel, 108 of them communicate solely through this channel). Hence, if such contacts would be inaccurately identified based on IM only, this could lead to a decline in accuracy and in turn decrease the supposedly higher average in accuracy when considering all channels and all identified contacts. Despite relying on different samples recruited several years apart, our results support the conclusion of [7, 8] studies.

5.4.3. Classification based on two classes: FAM vs. (ACQ ∪ FRI ∪ COL)

We consider FAM against the combination of all the other categories. Our results in Fig. 4 confirm that by using only two classes, the performance of Random Forest over $S_{15(IM)-SMOTE-2}$ dataset is further improved.

5.4.4. $S_{15-SMOTE}$ and $S_{25-SMOTE}$ performance variations.

We reached better results with $S_{15-SMOTE-4}$ and $S_{15-SMOTE-3}$ than their S_{25} counterparts, as shown in Fig. 3. To investigate this difference,

Table 10: Results obtained when removing a specific user (u16 to u25)

Dataset	Precision	Recall	F-measure
S25-overall	66.0	69.6	67.7
S25-u16	66.7	69.1	67.8
S25-u17	68.4	69.9	69.0
S25-u18	67.8	69.8	68.7
S25-u19	66.6	69.9	68.1
S25-u20	70.2	70.9	70.5
S25-u21	64.4	67.1	65.6
S25-u22	67.2	69.7	68.4
S25-u23	67.5	69.2	68.3
S25-u24	67.5	70.5	68.9
S25-u25	68.8	70.8	69.7

we analyze the individual impact of the ten participants, who are included in S_{25} but are missing in S_{15}, on the overall performance as follows: (1) We create ten different datasets, each omits one of these 10 users, (2) We run the SMOTE over-sampled technique and apply Random Forest. Tab. 10 shows that nine out of ten user-based measurements perform better when removed, showing that they actually have just a slightly negative impact on the entire dataset. Thus, the decrease in performance of $S_{25-SMOTE-4}$ as opposed to $S_{15-SMOTE-4}$ subsets can be attributed to them. In turn, this unexpected discrepancy between aforementioned subset results could be attributed to the variety of users' demographic backgrounds.

5.4.5. Feature Selection

We applied different feature selection techniques to select the most important features. The results shown in Tab. 11 demonstrate that the highest rated features are distributed between multiple channels, i.e, calls, IMs, and emails. Particularly, the features about early morning instant messages fea-

Table 11: Top ten features of $S_{15(IM)-SMOTE-4}$ derived from gain ratio feature evaluator using the Ranker search method [30].

Features	Gain ratio
Minimum number of calls per week during whole interval	0.2195
Number of early morning IM emojis	0.2107
Length of early morning IMs	0.2032
Number of early morning IMs	0.1996
Minimum number of emails per week	0.1922
Minimum number of calls per week over last half year	0.1901
Length of late night calls divided by total length	0.1654
Number of late night calls divided by total number of calls	0.1588
Average number of emails per week	0.1526
Maximal number of emails per week	0.1490

tures, i.e., number of early morning IM emojis, length and number of early morning IM, have a high impact. Length and number of late night calls are also among the top ten worthy features. In contrast, Tab. 12 shows that the most worthy features are dominated by call-based ones. Particularly, the calls during unusual hours tend to be more worthy, i.e., length and number of late night calls, of Sunday calls, and early night calls.

In summary, our results show that (1) Random Forest performs best, (2) merging *coworkers* and *work/schoolmates* categories into just *colleagues* may lead to better results, and (3) IMs increase the precision and recall rate for four and three classes. Our results further indicate that by classifying based on two classes, the inclusion of instant messages leads to a decline in precision and recall. Our approach leads however to higher accuracy as opposed to related works [8, 11] over a finer-grained classification of four categories. This is particularly attributed to inclusion of e-mails and outgoing IMs (not included in [11]), respectively incoming and outgoing IMs (omited in [8]).

Table 12: Top ten features of $S_{25-SMOTE-4}$ derived from gain ratio feature evaluator using the Ranker search method [30].

Features	Gain ratio
Minimum number of calls per week during last half year	0.2338
Minimum number of emails per week	0.1807
Length of late night calls divided by total length	0.1622
Length of Sunday calls divided by total length	0.1602
Number of late night calls divided by total number of calls	0.1443
Number of Sunday calls divided by total number	0.1416
Minimal number of calls per week during whole interval	0.1365
Length of early night calls divided by total length	0.1324
Number of early night calls divided by total number of calls	0.1270
Length of late night SMS divided by total length	0.1259

6. Discussion

6.1. Convenience sampling

The closing of our institution and the pandemic restrictions impacted the recruitment of participants outside our social network. However, our results are independent of existing social relationships with us. As often in qualitative studies, our sample may not be representative. We however took care to include participants with different demographic profiles. We further compared our results with those we obtained in 2014 [8] (vs. 2021 in our case) with a sample of 19 participants (vs. 25) between 23 and 29 (vs. 18 and 65) based in Germany (vs. Kosovo, Germany, and Switzerland) and obtained similar results (see Sec. 5.4).

6.2. Selection bias

Our participants could freely decide to participate. A self-selection bias may thus exist. They also agreed to the collection of their metadata. This

may suggest a less concerned attitude towards privacy than others, potentially impacting how they communicate with others.

6.3. Number of participants

Our sample of 25 participants is in-line with existing studies (e.g., eight in [11], 19 in [8], and 22 in [22]) and longitudinal field study over several weeks [8, 11]. We expect that a higher number of participants would increase the classifiers accuracy. However, the announced changes in the WhatsApp privacy policy [1] and the related news lead us to prepone our study, as we expected WhatsApp users to move to other messengers, such as Signal.

6.4. Android OS

Our study is based on an Android app. We hence recruited our sample accordingly. While differences in perceptions and attitudes of, e.g., Android vs. iOS users have been studied in particular contexts [31–33] or privacy and security in general [33], there exists no studies to the best of our knowledge that analyze differences in communication patterns of users of different OS.

6.5. Partial communication patterns

Our app does not cover all communication channels. For example, they may have used other devices (e.g., computers, landline phones), other applications (e.g., video conferences, games), or other accounts (e.g., professional e-mails). As in [8], the participants could decide to log the metadata from one e-mail address or more and one messenger or more. As a result, the obtained metadata may provide only a partial view of their communication patterns. Nevertheless, we hypothesize that having even more data could lead to a better classification of the different social categories.

Note that we did not take into account existing groups in the considered instant messengers, although we collected metadata about the exchanged messages within the groups. Due to the nature of accessibility service, in group communications, the group name and the participants ID can be identified but not other group members. To obtain the information, the participants should have manually indicated which contacts belonged to each group, thus further greatly increasing the requested efforts. Moreover, it would have requested a complex analysis of the IM content to determine if the intended recipients are the whole group or particular members.

7. Conclusion

Nowadays, individuals could hardly imagine how to live without the use of online social networks and messengers. Many providers often argue how certain techniques, e.g., end-to-end encryption, preserve the secrecy of users' exchanged content. Whereas such techniques contribute to protect users' privacy, metadata are often neglected. We have however shown how mobile communication metadata could expose interpersonal relationships between users' social circle. We have particularly considered IM metadata, in addition to the other communication channels, i.e., calls, SMS, and e-mails. Our results indicate that users' interpersonal relationship can be predicted with a f-measure score of 79.4% for a four-classes classification model, namely, *acquaintances, friends, family*, and *colleagues*. Moreover, we have shown that when *acquaintances* and *friends* are merged in one class, beside *family* and *colleagues*, the f-measure score reaches 86.7%. When considering only two classes, i.e., *family* on one side and all the other merged on the other side, the f-measure score is 89.6%. A feature importance analysis have shown

that IM-based features, such as the length and number of early morning IMs, as well as, number of early morning emojis, provide good insights for the classification. Minimum number of calls and length of late night calls are also worthy attributes in both of the datasets (i.e., $S_{15(IM)-SMOTE-4}$ and $S_{25-SMOTE-4}$) While our results confirm that a relatively small set of mobile communication metadata can reveal social ties, we will conduct a cross-cultural study and consider additional sources of information, such as metadata from uploaded stories, to extend these results.

Acknowledgment

We thank our participants and our students for their preliminary works. This project is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) referenced with the number #317687129.

References

[1] WhatsApp Privacy Policy, Online: `https://www.whatsapp.com/legal/updates/privacy-policy/` (accessed in 25.01.2022) (2021).

[2] J. Tang, H. Shoemaker, A. Lerner, E. Birrell, Defining Privacy: How Users Interpret Technical Terms in Privacy Policies, In PoPETs, 2021.

[3] P. Pietromonaco, N. Collins, Interpersonal Mechanisms Linking Close Relationships to Health, The American Psychologist, 2003.

[4] S. Saeb, E. Lattie, K. Kording, D. Mohr, Mobile Phone Detection of Semantic Location and Its Relationship to Depression and Anxiety, JMIR mHealth and uHealth.

[5] B. Erickson, Social Networks: The Value of Variety, Contexts, 2017.

[6] J. Natasha, T. Sara, A. Asaph, G. Asma, S. Akane, P. Rosalind, Predicting Students' Hapiness From Physiology, Phone, Mobility and Behavioral Data, Proc. of IEEE ACII, 2015.

[7] J.-K. Min, J. Wiese, J. I. Hong, J. Zimmerman, Mining Smartphone Data to Classify Life-facets of Social Relationships., Proc. of ACM CSCW, 2013.

[8] D. Reinhardt, F. Engelmann, A. Moerov, M. Hollick, Show Me Your Phone, I Will Tell You Who Your Friends Are: Analyzing Smartphone Data to Identify Social Relationships, Proc. of ACM MUM, 2015.

[9] Most Popular Messaging Apps (Around the Globe), Online: `https://www.statista.com/statistics/258749/most-popular-global-mobile-messenger-apps/` (accessed in 25.01.2022) (2021).

[10] M. Winik, Most Popular Global Mobile Messenger Apps as of April 2021 Based on Number of Monthly Active Users, Online: `https://www.similarweb.com/corp/blog/research/market-research/worldwide-messaging-apps/` (accessed in 25.01.2022) (2021).

[11] R. Dwarakanath, J. Charrier, F. Englert, R. Hans, D. Stingl, R. Steinmetz, Analyzing the Influence of Instant Messaging on User Relationship Estimation, Proc. of IEEE MS, 2016.

[12] D. Reinhardt, F. Engelmann, M. Hollick, Can I Help You Setting Your Privacy? A Survey-based Exploration of Users' Attitudes towards Privacy Suggestions, in: Proc. of ACM MoMM, 2015.

[13] L. Fang, K. LeFevre, Privacy Wizards for Social Networking Sites, in: Proc. of WWW, 2010.

[14] M. Johnson, S. Egelman, S. M. Bellovin, Facebook and Privacy: It's Complicated, in: Proc. of SOUPS, 2012.

[15] P. G. Kelley, R. Brewer, Y. Mayer, L. F. Cranor, N. Sadeh, An Investigation into Facebook Friend Grouping, in: Proc of IFIP INTERACT, 2011.

[16] Y. Wang, G. Norcie, S. Komanduri, A. Acquisti, P. G. Leon, L. F. Cranor, I Regretted the Minute I Pressed Share: A Qualitative Study of Regrets on Facebook, Proc. of SOUPS, 2011.

[17] J. Wiese, P. G. Kelley, L. F. Cranor, L. Dabbish, J. I. Hong, J. Zimmerman, Are You Close with Me? Are You Nearby?: Investigating Social Groups Closeness and Willingness to Share, Proc. of ACM UbiComp, 2011.

[18] M. Namara, H. Sloan, P. Jaiswal, B. P. Knijnenburg, The Potential for User-Tailored Privacy on Facebook, Proc. of IEEE PAC, 2018.

[19] M. Zeina, F. Yakoub, A. Adl, A. E. Hassanien, V. Snaselc, Identifying Circles of Relations from Smartphone Photo Gallery, Proc. of ICCMIT, 2015.

[20] H. Hsieh, C. Li., Inferring Social Relationships From Mobile Sensor Data, Proc. of WWW, 2014.

[21] T. Liu, J. Nicholas, M. M. Theilig, S. C. Guntuku, K. Kording, D. C. Mohr, L. Ungar, Machine Learning for Phone-Based Relationship Esti-

mation: The Need to Consider Population Heterogeneity, Proc. of ACM IMWUT, 2019.

[22] J. Choi, S. Heo, J. Han, G. Lee, J. Song, Mining Social Relationship Types in an Organization using Communication Patterns, Proc. of CSCW, 2013.

[23] X. Bao, J. Yang, Z. Yan, L. Luo, Y. Jiang, E. M. Tapia, E. Welbourne, CommSense: Identify Social Relationship with Phone Contacts via Mining Communications, Proc. of CSCW, 2013.

[24] S. H. Mirisaee, S. Noorzadeh, A. Sami, R. Sameni, Mining Friendship from Cell-Phone Switch Data, Proc. of IEEE HumanCom, 2010.

[25] D. Christin, A. Bentolila, M. Hollick, Friend is Calling: Exploiting Mobile Phone Data to Help Users in Setting their Privacy Preferences, Proc. of IWSSI/SPMU, 2012.

[26] J. Wiese, J. Min, J. I. Hong, J. Zimmerman, You Never Call, You Never Write: Call and SMS Logs Do Not Always Indicate Tie Strength, Proc. of ACM CSCW, 2015.

[27] J. Mayer, P. Mutchler, J. Mitchell, Evaluating the Privacy Properties Of Telephone Metadata, Proc. of the National Academy of Sciences, 2016.

[28] M. Brandt, Instant Messenger Market Share in Germany, Online: `https://de.statista.com/infografik/23449/umfrage-zur-nutzung-von-messengern-in-deutschland/` (accessed in 25.01.2022) (2021).

[29] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, SMOTE: Synthetic Minority Over-Sampling Technique, Journal of Artificial Intelligence Research, 2012.

[30] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten, The WEKA Data Mining Software: An Update, ACM Special Interest Group on Knowledge Discovery in Data Explorations Newsletter, 2009.

[31] H. K. Ubhi, D. Kotz, S. Michie, O. C. P. V. Schayck, R. West, A Comparison of the Characteristics of iOS and Android Users of a Smoking Cessation App, Journal of Translational Behavioral Medicine (2017).

[32] R. Pryss, M. Reichert, W. Schlee, M. Spiliopoulou, B. Langguth, T. Probst, Differences Between Android and iOS Users of the Trackyourtinnitus Mobile Crowdsensing Mhealth Platform, Proc. of IEEE CBMS, 2018.

[33] P. Ünal, T. T. Temizel, P. E. Eren, A Study on User Perception of Mobile Commerce for Android and Ios Device Users, in: Proc. of MobiWis, 2015.