App-generated digital identities extracted through Android permission-based data access - a survey of app privacy

Nurul Momen[1] Lothar Fritsch [2]

Abstract: Smartphone apps that run on Android devices can access many types of personal information. Such information can be used to identify, profile and track the device users when mapped into digital identity attributes. This article presents a model of identifiability through access to personal data protected by the Android access control mechanism called *permissions*. We populate partial identities with attributes related to permission-protected personal data, and then show how apps accumulate such attributes in a longitudinal study that was carried out over several months. We found that apps' successive access to permissions accumulates such identity attributes, where apps show different interest in such attributes.

Keywords: Privacy; Android; Apps; Identification; Digital Identity; Survey and Permissions

1 Introduction and research question

Smartphones and other small communication devices using the Android operating system are tools inseparable from everyday lives for many human beings. Most aspects of communication, interaction and organization of their lives are channeled through smartphone-based apps. These devices accumulate and concentrate large amounts of personal data in contact lists, digital calendars, phone call logs, text message archives, GPS-labeled photo collections and other data. In addition the smartphone hardware offers sensing and intelligence gathering capabilities that allow data collection such as:

Frequent precise positioning through GPS or the collection of IP addresses and GSM cell information,

Collecting inventories of electronic equipment through scanning of the wireless and Bluetooth environment including own and neighbor home electronics and other person's smartphones,

Tracking phone appearance in contexts such as supermarkets, trains, airports or restaurants by mapping their Wifi hot-spots,

[1] Karlstad university, Dept. of Mathematics and Computer Science, Universitetsgatan 2, Karlstad, Sweden nurul.momen@kau.se
[2] Karlstad university, Dept. of Mathematics and Computer Science, Universitetsgatan 2, Karlstad, Sweden lothar.fritsch@kau.se

doi:10.18420/sicherheit2020_01

Acquisition of biometric and behavioural patterns through location, motion sensors, biometric sensors, cameras and microphones,

Profile media use through audio beacons in music streams and TV programs.

Such intelligence collection accumulates identifying information. The research presented in this article is concerned with how such information informs identity attributes following the model of partial identities proposed by Pfitzmann and Hansen [PH10].

Since digital identifiers can create large profiles of private behaviour when they get used and observed by third parties, the accumulation of information about partial identity attributes can cause privacy problems [PF11]. However, the exact level of how identifiable smartphone users are through such identity profile remains mostly unknown to them, since app-based information extraction and accumulation remains mainly opaque [Mo17]. Recent research on identifiability in anonymized or pseudonymized data collections shows that fifteen or less anonymized attributes are sufficient to re-identify a person in a database [RHDM19].

To investigate these matters, we populate partial identity attributes with statistically aggregated data from long-term monitoring of apps on Android platforms [Mo18a] through a study of app permission access.

Research questions:

1. *Which identity attributes can apps extract through accessing permission-protected smartphone data?* Our research aims at mapping smartphone data into identifiable attributes to show the overall collection of identity attributes extracted per app.
2. *What permission-protected identity attribute data are the apps accessing the most?* Through analysis of a longitudinal sample of app permission use we investigate the aggregated long-term identity attribute profiles gathered about smartphone users.

Outline: The rest of this article is organized as follows: First, we explain our model in Section 2 and the data acquisition and survey method for identity attribute extraction of mobile apps in Section 3. Section 4 describes our results from profiling the identity attribute extraction of 50 apps. Finally, we conclude this paper and point out directions for future research in Section 5.

2 A model for permission-based partial identity

This section will introduce our model of generated, identifiable digital identities extracted from smartphones. As shown in [FM17], there is a correlation between partial identity attributes and access control logs from Android app permission. It defines eight partial identity attributes which comprise the smartphone identity of an individual. Though only

ten permission groups are defined as *dangerous* by Android, there are more than 300 permissions listed in the master branch of its' code base which have the potential to yield identifiable information [Go19]. This section elaborates on the background of our work. Unlike previous work which focused on static analysis of app code[Ha18] or on traffic analysis[Ma12] to measure app behavior, we record permission access of running apps over a period of time.

2.1 Personal identification in digital data

Digital identity is often defined as an identifier with related identity attributes attached [Cl09]. Pfitzmann and Hansen [PH10] defined: *An identity is any subset of attribute values of an individual person which sufficiently identifies this individual person within any set of persons.* Note that seldom there is a single identity for a person, but there exist many combinations and permutations of identity attributes that are used in various sets (relating to distinct social contexts or situations they get applied to). Therefore, they introduce the concept of a partial identity by defining: *A partial identity is a subset of attribute values of a complete identity, where a complete identity is the union of all attribute values of all identities of this person.* A person is identifiable through identity attributes if s/he is easily linked to a digital identity. A person is then supposed to be linkable if he or she can get identified in a data set based on the combination of partial identity attributes used for the transaction. The concept of the partial identity is further used to define relationships between identity and attribute data as well as relationships between sets of attributes. The authors of [PH10] define anonymity, unlinkability and unobservability properties based on the concept of partial identities. One important observation is that a person is unidentifiable (anonymous) in a data set if that person's attributes cannot get identified within that data set. On the other hand, a person may become more identifiable, once more attributes get added to a partial identity. So, identifiability is directly proportional to accumulated partial identity information. Based on the concept of partial identities, we accumulate information gained through app permissions for the partial identity sets that can get retrieved through the permissions. In the next sections, a model is described for building partial identities from information accessible through permissions on Android devices and an empirical study is presented indicating the likelihood of partial identity extraction.

2.2 Partial Identities

For this survey, we use an improved version of the model for partial identities published in [FM17]. The model now leaves out the derived identity attributes. It contains additional edges mapping permission groups into identity attributes. The reduction to direct attributes serves the purpose of computability of accessed identities in the context of our survey. Analyzing the information accessible through those permissions, we mapped identity attribute building information sources to identity attributes that get collected from the

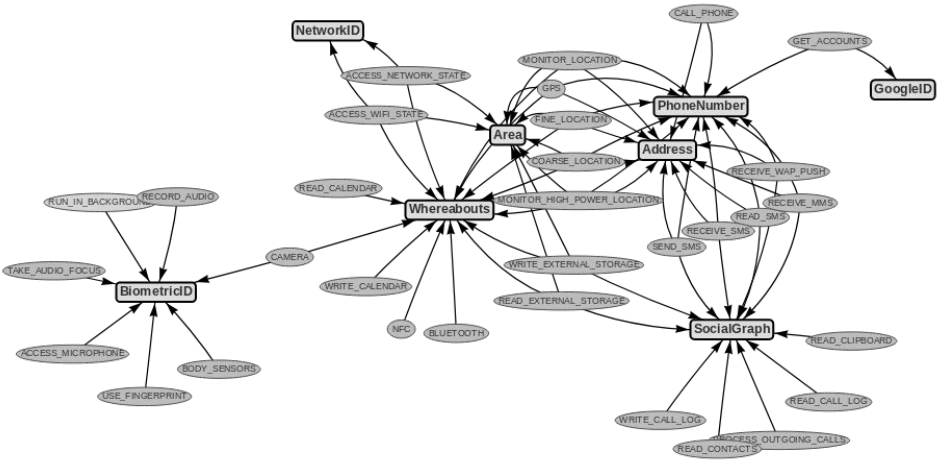

Fig. 1: Permissions (elliptical shapes) contribute to observable partial identities (rectangular shapes), expressed by directed edges.

information sources. From the permissions perspective, the partial identity P of an Android smartphone user is defined as:

$$P = \{Whereabouts, NetworkID, GoogleID, BiometricID, PhoneNumber, Address, Area, SocialGraph\}$$

It should be noted that there are directly accessible identity attributes, such as phone numbers or fine-grained GPS location, as well as data that can be used to derive identity attributes through various techniques. We defined three distinct attributes that are related to location, since the quality of their identity attribute extraction is quite different: *Whereabouts, Area, Address*. The following list defines the directly accessible partial identity attributes, lists the data contributing permissions. Permission groups are sorted into partial identity attributes they inform when extracted from a smartphone.

SocialGraph : *Gathers social graph elements from* READ_CONTACTS, READ_CLIPBOARD, CALL_LOG, EXTERNAL_STORAGE

Address : *Retrieves address from* SMS, CALL_PHONE, LOCATION

PhoneNumber : *Retrieves phone number from* CALL_PHONE, GET_ACCOUNTS, LOCATION, SMS

GoogleID : *Get Google ID from* GET_ACCOUNTS

Whereabouts : *Precise location from* LOCATION, EXTERNAL_STORAGE, NFC, CAMERA, BLUETOOTH, CALENDAR, ACCESS_WIFI_STATE, ACCESS_NETWORK_STATE

NetworkID : *Network access ID from* ACCESS_WIFI_STATE, ACCESS_NETWORK_STATE

Area : *Retrieves geographic area from* LOCATION, EXTERNAL_STORAGE, ACCESS_WIFI_STATE, ACCESS_NETWORK_STATE

BiometricID : *Collects biometric information from* CAMERA, USE_FINGERPRINT, RECORD_AUDIO, RUN_IN_BACKGROUND, BODY_SENSOR

As shown in Fig. 1, the relationship between accessed permissions and partial identity attributes can be visualized as a directed graph where the partial identities and permissions are nodes that are connected with unidirectional edges [FM17].

3 Methodology and data collection

This section describes our method and the implementation of the data collection for survey. Figure 2 shows the overall procedure. Based on popularity, we installed a sample of the 50 most popular apps from five categories: *Communication*, *Social*, *Fitness*, *Weather*, and *Music*. These apps were installed on Nokia 5 smartphones with Android 7.1.1 (Nougat). They had the pre-configured monitoring tool *Aware* installed that collected apps' permission access logs [Mo18b]. Then the partial identity model was applied to the collected data.

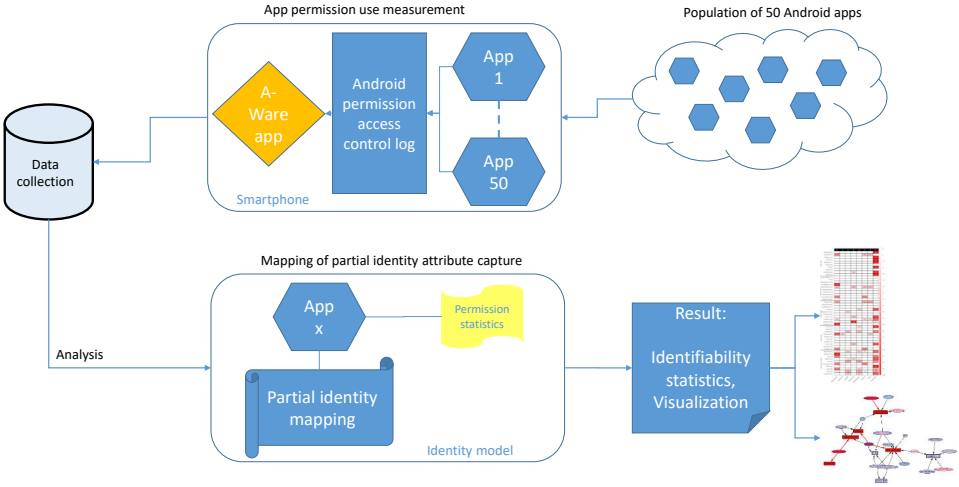

Fig. 2: Data collection: Permission logging using *Aware*, followed by mapping into partial identity model.

Capturing permission-based access control data: Accessibility to user data through permissions gives full discretionary access for the app without any constraints. We measure apps' permission access patterns using the *Aware*-client based on the method described in [Mo18b]. In brief, each permission request delivered to the Android system will be logged with date, time and location, together with the app package name. The experimental setup is described in detail in [Ha19]. *Aware* was run 24/7 on seven smartphones over the data collection period. Apps and categories considered in this article are shown in Figure 4.

Data collection: The app permission access patterns were gathered from seven Android smartphones in the time period 11 December 2018 to 11 March 2019. 50 popular Play store apps in 5 different app categories were installed. All installed apps were started, personalized and then left alone while the phones were resting or were taken to various locations. Apps were running in background at their own discretion. The smartphones had mobile 4G data access as well as configured Wifi connections to ensure permanent connectivity. In total, 664339 lines of log were accumulated in the survey period. The accumulated log data is structured as shown in Table 1.

package_name	permission_name	timestamp
com.whatsapp	READ_EXTERNAL_STORAGE	Mon Mar 11 10:36:50 GMT+01:00 2019
com.joelapenna.foursquared	MONITOR_LOCATION	Sun Mar 10 19:28:20 GMT+01:00 2019
com.google.android.apps.fitness	READ_EXTERNAL_STORAGE	Tue Feb 19 19:25:39 GMT+01:00 2019
com.yahoo.mobile.client.android.weather	MONITOR_HIGH_POWER_LOCATION	Tue Jan 15 10:02:02 GMT+01:00 2019
com.spotify.music	TAKE_AUDIO_FOCUS	Fri Mar 08 19:13:17 GMT+01:00 2019

Tab. 1: Sample of data captured from Android permission access control. Shown: App package name, permission accessed, timestamp.

Mapping the permission access data into the model of partial identities: Data selection and pre-processing were performed with the *KAUDroid* data visualization tool [SBB19] that was implemented by students at Karlstad University. The tool supports import and selection of *Aware* log files from the KAUDroid database [Ca18], the selection of apps and date ranges as well as permission groups for visualization. It implements various views on the selected data, one of which is the graph visualization of partial identity extraction based on permissions. Figure 3 shows the case of *Whatsapp*: partial identity model is applied to the permission access logs which contribute to formulate five partial identity attributes out of eight. From the collected log for *Whatsapp*, 1834 entries contributed to attribute formulation out of 3770 entries.

4 Results

Our accumulated data in Figure 4 shows large differences between apps. The figure shows in each row the percentage of accessed identity attributes per app. Numbers and coloring

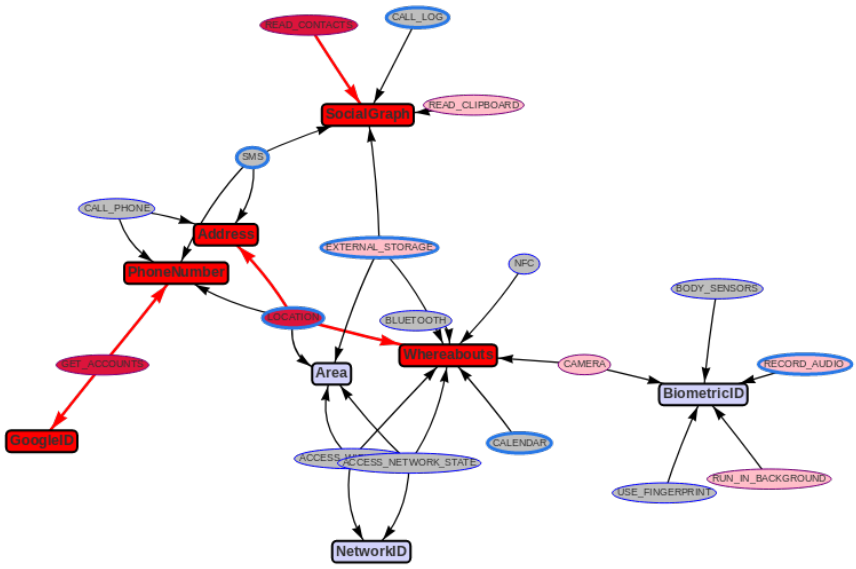

Fig. 3: Partial identity model applied on *Whatsapp*'s permission use. The extracted identity attributes are colored in intense red color, while the informing permissions linked to them that were measured are colored with light red color.*Whatsapp* extracts five identity attributes.

indicate the weighting of access. The numbers show each attributes percentage of the total attribute accesses. Each row has a total of 100 per cent. App *Pedometer* - a training tracking app - for example extracts mainly the partial identity attributes *BiometricID* (95%) and *SocialGraph* (5%), while App *Viber* - a communication and messaging app - extracts five identity attributes with a focus on *SocialGraph* (33%), *PhoneNumber* (29%) and *GoogleID* (29%).

Social relationships and location profiling: Majority of the apps extract social graph attribute related information and whereabouts (location information with high degree of detail). This is visible in columns A and H of Figure 4. We observe therefore that detailed geographic information and the social graph of a smartphone user is the most sought-after combination of identity attributes.

***Communication* and *social* apps extract most attributes:** The two app groups that show the widest extraction of identity attributes are the communication apps and the Fitness apps, with the latter group extracting half of the attributes for the firstly mentioned group. Table 2 shows the average attributes collected per group - and the number of apps with broad attribute access (4 or more).

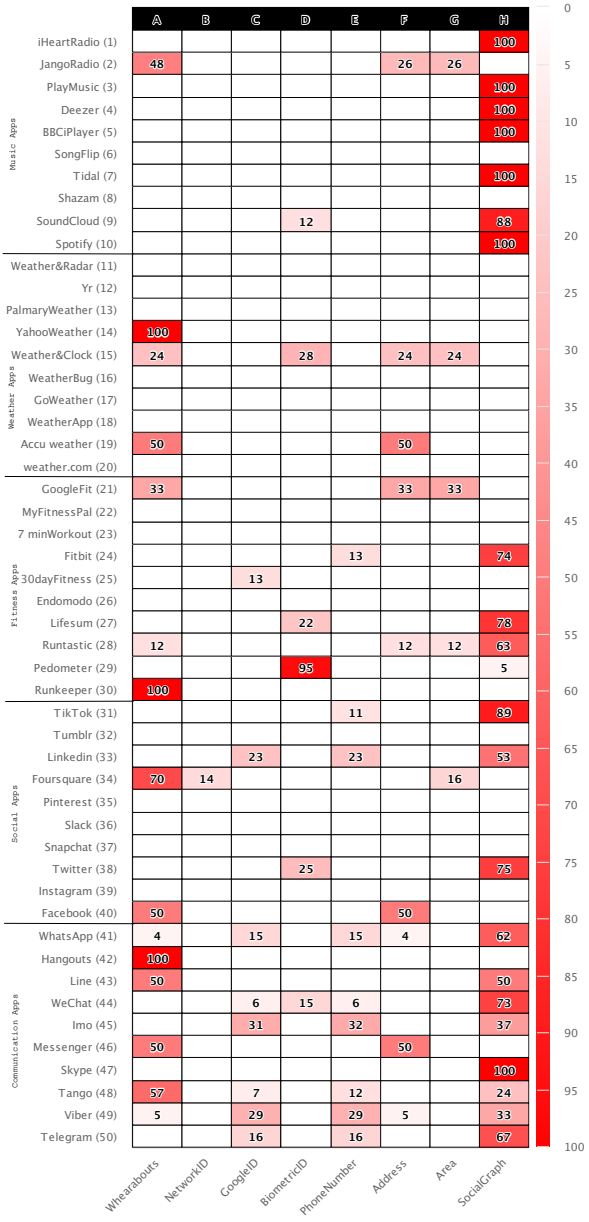

Fig. 4: Identifiability per app, as of partial identity collected. Numbers shown are percentage of access to identity attribute compared to all accesses.

	Music	**Weather**	**Fitness**	**Social**	**Communication**
Average count of attrib.	1.1	0.7	1.5	1.2	3.0
Apps using 4+ attributes	0	1	1	0	4

Tab. 2: Average number of ID attributes extracted per group and number of apps that access four or more attributes.

Upon examination of individual apps, we noticed that there are several apps that access identity attributes that may not be necessary to fulfil the advertised purpose of the respective app. *Weather&Clock* is accessing all location-related attributes, even the high-resolution ones that are much too detailed for weather forecast apps. In addition the app is interested in biometric identity attributes. A second example is *Runtastic* in the *Fitness* category which extracts *SocialGraph* attributes besides all location attributes.

Comparing Figure 4 to a former analysis of app permission showing we notice consistency of our results. In Figure 3b in [MHF19] we show that app permission access is broadest and most frequent in the communication app group. The Fitness apps and social apps are shown to be very active, too. The potential privacy risk of *Runtastic* has been noticed in a previous study as well [Ha19].

A human-friendly visual presentation: To facilitate visual analysis, we show a graph view of user identifiability through apps. The diagram in Figure 3 shows access to five partial identity attributes for *Whatsapp*. Further visualisations are provided in the appendix. The graphs are created using the *KAUDroid* tool [SBB19]. The graph view may facilitate various intuitive views. By coloring extracted attributes, it provides an overview of how identifiable a user is as seen by an app. We plan to add the magnitude of identification to the nodes such that it will be intuitive to distinguish between apps that access attributes more often and those that access attributes less.

We have to point out that our sample of 50 apps may not be representative for the larger population of apps. However we see that our analysis through partial identity profiling provides an additional perspective on app privacy issues.

5 Discussion and conclusion

Our survey shows major differences in the number of identity attributes extracted by apps. The group extracting the highest number of identity attributes are the *Communication apps*, followed by the *Fitness apps*. Concerning research question 1, we provided and tested a mapping of access to groups of Android permissions into a model of identity attributes. By collecting permission showing data, and by mapping it into our model, we showed in Figure 4 that each of our attributes is accessed by apps. We therefore constitute that we found a mapping of permission-protected data on identity attributes that works. Research question 2 provided an overview of apps that are overly concerned with collecting identity attributes

that relate to a smartphone user's social network and that provide the phone's detailed whereabouts. The other attributes in our model are extracted at much smaller numbers. We therefore answered this question by finding that social relations and location are the major identifiable attributes apps are interested in. Our findings are puzzling at first, however the social graph and detailed location profiles over time will identify a person sufficiently. Home and work locations as well as communication patterns with social contacts will not change quickly for most human beings. We must therefore expect that the re-identification potential based on a partial identity $P = \{Whereabouts, SocialGraph\}$ is sufficient for most tracking applications when used with forensic tools (see [MHW12] and [MHS11] for insights in the forensic potential of social networks and smartphone location tracks, respectively).

Utility of results and future work: Our model shows that we can rank apps by how identifiable their users are through partial identities. We see three major applications for our results. First, such rankings may help consumers make decisions about which app they will use to solve a task. Then, app rankings may constitute useful information for regulatory authorities who oversee data protection legislation and who need data about actual app behavior when data is sourced e.g. through citizen science crowdsourcing [BZNT11]. Finally, through graphical visualization, our model can provide insight for users when used on their individual data. As one path for future work we plan to combine the *Aware* data capture app with the partial identity model and the graph visualizations into a personal transparency-enhancing tool (TET) [MFH17] for end users. The tool will be useful to evaluate own exposure and identifiability against app providers. Further improvement of the data aggregation that populates the model will be necessary. As of now we equally weigh all permissions and all identity attributes independent of their contribution to identification risk and independent of their potential differences in privacy impact. We consider personalized weights as a mechanism that will allow the selection of privacy personas as default calibration of the data analysis.

Some challenges of our approach: Apps are not a static infrastructure. They constantly update themselves. Their actual behavior is triggered by their user, by their service provider pushing messages to apps, and in the case of social media the behavior is based on the social network member's activities, too. Such dynamics create difficulties for reproducible survey results. Many of those difficulties are described in Section 1.3 in [MHF19]. One major challenge is the reproduction of user behaviour including the separation of experimental use and private activities on smartphones [Go17]. To avoid major difficulties, we decided to install the apps, to personalize them by creating profiles or accounts, and then leave the devices undisturbed on our desks. The resulting data shows app behavior based on idle-time app activity. Our results are therefore presumably a subset of potential app activity with regular interaction. Additional challenges are the frequent updates of Android and especially of the permission system.

Summarizing our findings, we successfully mapped app's access to smartphone data to a model of digital partial identity which we then used to show the differences and specific

behaviors in a sample of 50 popular consumer apps. In particular, the graphical visualization can be used to inform about identifiability risk through such apps.

Acknowledgements

The research leading to this publication was partially funded by the Norwegian Research Council ArsForensica project nr.248094/O70.

References

[BZNT11] Burguera, Iker; Zurutuza, Urko; Nadjm-Tehrani, Simin: Crowdroid: behavior-based malware detection system for android. In: Proceedings of the 1st ACM workshop on Security and privacy in smartphones and mobile devices. ACM, pp. 15–26, 2011.

[Ca18] Carlsson, Adrian; Pedersen, Christian; Persson, Fredrik; Söderlund, Gustaf: KAUDroid : A tool that will spy on applications and how they spy on their users. report, Karlstad University, 2018. Project report.

[Cl09] Clarke, Roger: A sufficiently rich model of (id) entity, authentication and authorisation. In: The 2nd Multidisciplinary Workshop on Identity in the Information Society, LSE. volume 5, 2009.

[FM17] Fritsch, Lothar; Momen, Nurul: Derived Partial Identities Generated from App Permissions. In: Proceedings of the Open Identity Summit (OID) 2017, Lecture Notes in Informatics LNI, 277. Gesellschaft für Informatik, 2017.

[Go17] Goodyear, Victoria A.: Social media, apps and wearable technologies: navigating ethical dilemmas and procedures. Qualitative Research in Sport, Exercise and Health, 9(3):285–302, 2017.

[Go19] Google Git: Android Master Branch: , Number of permissions. `https://android.googlesource.com/platform/frameworks/base/+/master/core/res/AndroidManifest.xml`, 2019. Accessed on 14-Oct-2019.

[Ha18] Habib, Sheikh Mahbub; Alexopoulos, Nikolaos; Islam, Md Monirul; Heider, Jens; Marsh, Stephen; Müehlhäeuser, Max: Trust4App: automating trustworthiness assessment of mobile applications. In: 2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE). IEEE, pp. 124–135, 2018.

[Ha19] Hatamian, Majid; Momen, Nurul; Fritsch, Lothar; Rannenberg, Kai: A Multilateral Privacy Impact Analysis Method for Android Apps. In (Naldi, Maurizio; Italiano, Giuseppe F.; Rannenberg, Kai; Medina, Manel; Bourka, Athena, eds): Privacy Technologies and Policy. Springer International Publishing, Cham, pp. 87–106, 2019.

[Ma12] Marforio, Claudio; Ritzdorf, Hubert; Francillon, Aurélien; Capkun, Srdjan: Analysis of the communication between colluding applications on modern smartphones. In: Proceedings of the 28th Annual Computer Security Applications Conference. ACM, pp. 51–60, 2012.

[MFH17] Murmann, Patrick; Fischer-Hübner, Simone: Tools for Achieving Usable Ex Post Transparency: A Survey. IEEE Access, 5:22965–22991, 2017.

[MHF19] Momen, Nurul; Hatamian, Majid; Fritsch, Lothar: Did App Privacy Improve After the GDPR? IEEE Security & Privacy, 17(6):10–20, November-December 2019.

[MHS11] Maus, Stefan; Höfken, Hans; Schuba, Marko: Forensic analysis of geodata in android smartphones. In: International Conference on Cybercrime, Security and Digital Forensics, http://www. schuba. fh-aachen. de/papers/11-cyberforensics. pdf. 2011.

[MHW12] Mulazzani, Martin; Huber, Markus; Weippl, Edgar: Social network forensics: Tapping the data pool of social networks. In: Eighth Annual IFIP WG. volume 11. Citeseer, pp. 1–20, 2012.

[Mo17] Momen, Nurul; Pulls, Tobias; Fritsch, Lothar; Lindskog, Stefan: How Much Privilege Does an App Need? Investigating Resource Usage of Android Apps (Short Paper). In: 2017 15th Annual Conference on Privacy, Security and Trust (PST). pp. 268–2685, Aug 2017.

[Mo18a] Momen, Nurul: Towards Measuring Apps' Privacy-Friendliness (licentiate dissertation). Technical Report 2018:31, Karlstad University, Department of Mathematics and Computer Science, 2018.

[Mo18b] Momen, Nurul: Towards Measuring Apps' Privacy-Friendliness (licentiate thesis). PhD thesis, Karlstads universitet, 2018.

[PF11] Paintsil, Ebenezer; Fritsch, Lothar: A Taxonomy of Privacy and Security Risks Contributing Factors. In (Fischer-Hübner, Simone; Duquenoy, Penny; Hansen, Marit; Leenes, Ronald; Zhang, Ge, eds): Privacy and Identity Management for Life. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 52–63, 2011.

[PH10] Pfitzmann, Andreas; Hansen, Marit: Anonymity, unlinkability, unobservability, pseudonymity, and identity management-a consolidated proposal for terminology. In: Designing privacy enhancing technologies. Technische Universität Dresden, pp. 1–9, 10-Aug-2010.

[RHDM19] Rocher, Luc; Hendrickx, Julien M; De Montjoye, Yves-Alexandre: Estimating the success of re-identifications in incomplete datasets using generative models. Nature communications, 10(1):1–9, 2019.

[SBB19] Sundberg, Simon; Blomqvist, Alexander; Bromander, Anton: KAUdroid-Project Report: Visualizing how Android apps utilize permissions. report, Karlstad University, 2019.

Appendix A - Partial identity graphs

We show for further illustration the partial identity graphs for five selected app groups. The graphs show the app that extracts most identity attributes and the app that extracts the least attributes in the app groups. The graphs were created using the KAUDroid tool by Sundberg, Blomqvist, Bromander [SBB19].

The extracted identity attributes are colored in intense red color (dark), while the informing permissions linked to them that were measured are colored with light red color.

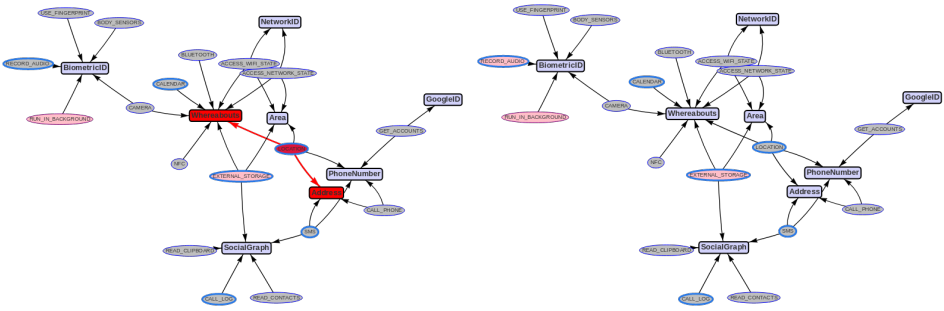

Fig. 5: JangoRadio and Shazam in *Music Apps* are the most and least user identifying apps.

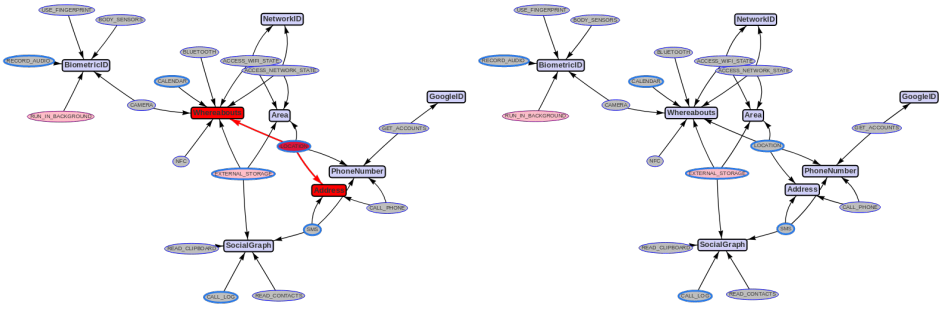

Fig. 6: Weather&Clock and PalmaryWeather in *Weather Apps* are the most and least user identifying apps.

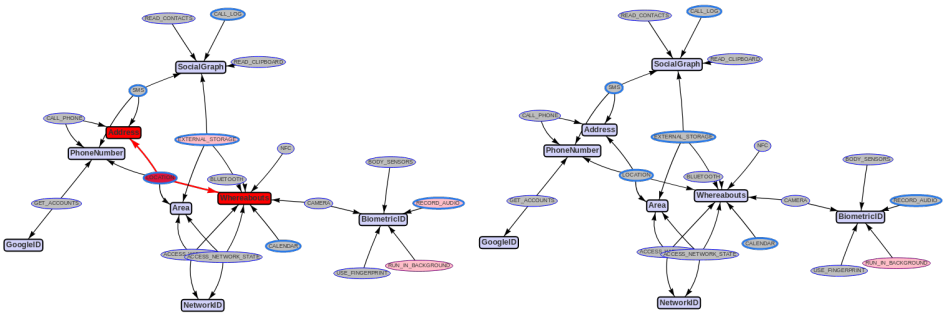

Fig. 7: Runkeeper and Endomondo in *Fitness Apps* are the most and least user identifying apps.

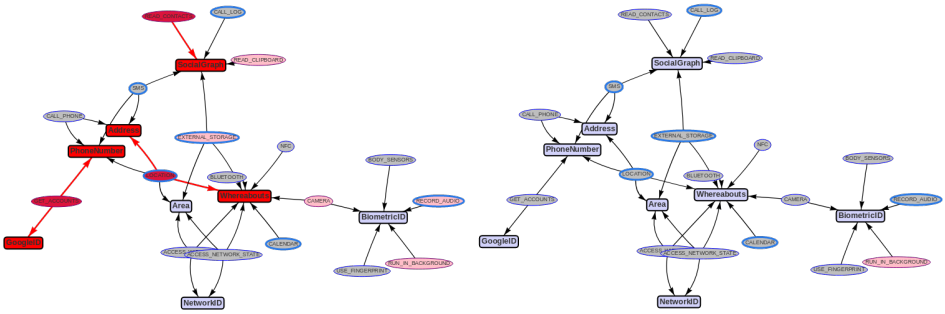

Fig. 8: Whatsapp and Instagram in *Social Apps* are the most and least user identifying apps.

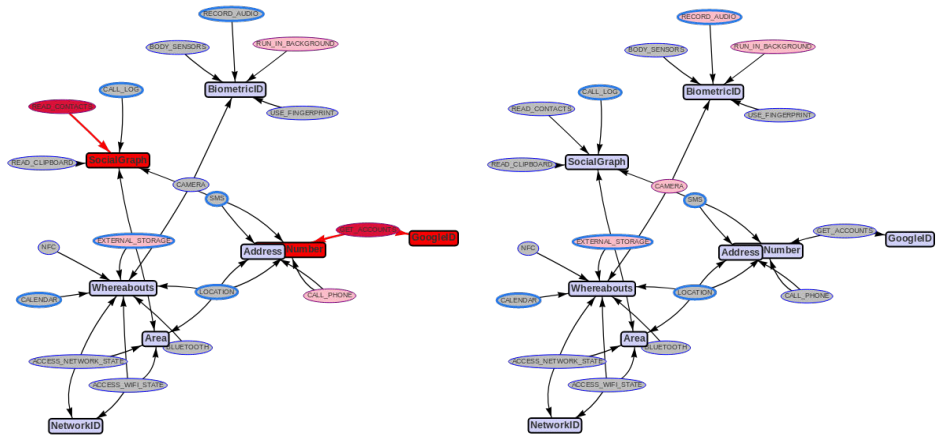

Fig. 9: Telegram and Skype in *Communication Apps* are the most and least user identifying apps.