Re-identification risk in mobile phone data

**Keywords:** Mobile phone data, CDR, privacy protection, privacy risk assessment.

# Introduction

The use of mobile phone data (MPD) for official statistical purpose is increasing. MPD are very useful for statistics related to population, migration, mobility. For instance, in developing countries, MPD can provide update estimates of population density, in the absence of other sources, see [1] for a review; in developed countries, often population registers are available, however, MPD provide more timely and detailed information, describing habits and behaviors that are not reported in the registers but they are also important for policy-making etc., e.g. human mobility, population density at a given time-space. Moreover, MPD can be used as auxiliary information for topics like poverty, SDG indicators, [1], [2] and [3] . Finally, MPD allow evaluating the coverage of the population registers [4].

However, the utility of MDP should be balanced by the risk for privacy violation of personal data. In fact, even if MPD are provided without direct identifiers (e.g. name, surname, date of birth, address, personal tax code) we cannot state they are anonymous. Several works claimed that it is possible to isolate a subject in a MPD database or link the MPD to subjects in different databases, or deduce, with significant probability, a characteristic of a subject from the MPD, [5] and [6]. So, MPD should be considered as personal data according to the GDPR [7], with an evaluation of the risk of re-identifying a person, even if personal data has been de-identified, encrypted or pseudonymised. Hence, to allow using MPD in a privacy preserving framework, a data protection impact assessment should be evaluated. This means to describe the planned processing operations, to assess the risks to privacy, to plan the measures to address those risks.

In this work, we investigate the risk of privacy and provide statistical measures. In cooperation with mobile phone provider, we apply our investigation to real data.

# Data description

The Call Details Records, CDRs, are pieces of information automatically stored by mobile operators to billing activities and to monitor the hand-over between network cells. Any active use of a mobile phone (call and text message in and out, in the following we shortly say call) is traced by a CDR. Under a framework agreement for research purposes in official statistics, the mobile operator shared with a NSI the following information about CDRs: a unique ID for each calling phone, time when the call starts, call duration, cell ID where the call starts and cell ID where the call ends. The information of the cell ID is referred to geographical coordinates of the tower antenna and the antenna technology. The data is shared without direct identifiers, and the table linking them to the supplier's archive is destroyed.

A crucial role is played by the device position, that is the location of the mobile phone when it calls. There are different frameworks for positioning, for instance handset-based, network-based or GPS-based. The choice depends on the type of data processed and the different network standards (GSM, CDMA, 3G), the purpose of the analysis of positioning information. Moreover, the location data (for example the location data function on a mobile phone) can be considered a personal data according to the GDPR.

To assess the mobile position, the characteristics of the mobile phone network cannot be disregarded: the cellular network is based on a set of base stations, usually consisting of a tower and several directional antennae. The radio coverage of a single antenna forms a network cell; several antennae form a cellular network. The size of a network cell and all cellular networks is not fixed; the phone normally switches to the closest antenna. However, if the network is crowded or visibility is disturbed, the phones can be switched not to the nearest station but to any other in the neighborhood. Figure 1 shows an example of antenna tower, cells and their coverage.



Figure 1. Example of antenna tower, cell and covered areas from mobile network

The maximum distance between a phone and an antenna depends on the antenna technology and the network dimensioning, and can vary from a few dozen to hundreds of kilometres.

# Methods

In this work, we focalize on the usage of MPD in an NSI and on the privacy attacks and privacy risks that are likely to occur even when MPD are provided without identifiers. In particular, we devote our attention to the case in which the external knowledge comes from statistical population registers and employees-employers databases, and these can be compared to the MPD in order to identify a single user. Hence, we consider a privacy attack, defined as “Home and Work Attack” where an intruder knows the two most frequent locations of an individual and their frequencies. As in [8] and [9] we define the privacy risk, or re-identification risk as the maximum probability of re-identification in a given attack scenario, and, given an attack, i.e. given some external knowledge *k*, the probability of re-identification for an individual *i* is given by:

$$PI=\frac{1}{\sum\_{k}^{}I\_{i}(k=i)} $$

where $I\_{i}(k=i)$ is the indicator function of the individual *i* matching the external knowledge *k.* The *PI*=0 if there is no record *i* matching external info *k*.

A similar concept is used in [6], where the privacy risk is related to the number of record/trajectories which are uniquely identified. Indeed, in [6] the privacy risk in CDR is studied through the analysis of mobility traces. A CDR record allows us to locate a user at a specific time, and a list of record of the same user draws a trace of how the user moves over time. The thesis is that even if the users are anonymized these can be re-identified by their traces of spatio-temporal points. In fact, if only one user reaches all the points of the trace (the trace is unique), this can be identified by an external source. Using real MDP, [6] prove that four spatio-temporal points are enough to uniquely identify 95% of users. Furthermore, by decreasing the spatial and temporal resolution of data, the power of identification decays very slowly.

# Experimental results on real data

To analyse the risk of privacy violation, we consider the CDRs from a mobile operator observed for six weeks in an administrative area of a medium-size (about 400000 inhabitants). In our experiment, we only consider users with at least three calls. We observe, on average, 47 calls per user per month. This value is significantly lower than the 114 contacts/month declared in [6]. This is due, in addition to the different behaviors of people, to the fact that our CDR contains only records for outgoing contacts (calls and text messages) while in [6] dataset contains all interactions of user with his phone.

We study some scenarios to measure the uniqueness of mobility trace in our CDRs. Let us first consider the “home and work” scenario described in previous section. We assign as home antenna the one with the highest number of calls during the nighttime (6pm-8am); the work antenna is assigned by the antenna with the highest number of calls during the daytime (9am-5pm) of the weekdays.

In this scenario, we observe that: 1.2% of traces is unique (i.e. just one user with home and work in that antennae), 1.1% of traces is twofold (exactly two users with that home and work antennae), 97.7% of traces can be considered as not at privacy risk. This analysis is quite complex to be rescaled at cell level, because several cells are related to each antenna tower and their dimensions are wide and overlapping.

Moreover, we consider the trace as a series of four spatio-temporal points (as proposed in [6]) where the antenna represents the spatial resolution and day and hour represents the time resolution. Analysing a huge sample of data, we observe that: 87% of traces is unique, 7% of traces is twofold, 6% of traces is covered by more than three users, so it can be considered as not at privacy risk. These results are very similar to the outcomes in [6]. The authors in [6] claim that when mobility trace are highly unique, the re-identification is easy using little outside information. However, from other sources it is likely to know the position of an individual (even in 4 points), but it is not so easy to know that the individual made a call. Indeed, from our data we can evaluate that the propensity to call is low: even in the peak hours, only one user over 12 makes a call. So, the probability that a user is calling in 4 points is negligible (less than 5×10-5).

# Conclusions

In the proposed work we describe two scenarios of privacy attack. In particular, we define likely scenarios when administrative data sources or other statistical information are available. Several issues remain open, the first one is related to the data localization via CDRs: an example of ambiguous traces descripted via CDRs is shown in figure 2. Other localization techniques, as GPS and sensors, could provide more punctual results, even in terms of both geographical location and frequencies of the device presence.

Another relevant point in assessing the privacy risk is related to the intruder’s knowledge of the mobile provider of each individual. Moreover, mobile phone provide information on human behaviours with timeliness compared to traditional sources, also the time lag can affect the effectiveness of a privacy attack.



Figure 2. Example of traces, antennas, cells and covered areas

In this work, we provide measures for the privacy risk, preparing actions to allow the usages of MPD in official statistics in a privacy-protected environment. Risk assessment is one of the fundamental elements to define the processing of data and the integration with security policies and privacy protection: this is the newness of the principle of privacy by design.

# References

1. UN Global Working Group on Big Data for Official Statistics (2017)Handbook on the use of Mobile Phone data for Official Statistics
2. Douglas, Meyer, Ram, Rideout, Song (2015) High resolution population estimates from telecommunications data, EPJ data science
3. Deville, Linard, Martin, Gilbert, Stevend, Gaughan, Blondel, Tatem (2014) Dynamic population mapping using mobile phone data, PNAS, vol. 111, no.45
4. De Fausti, Fabbri, Radini, Savarese, Spada, Tuoto and Valentino (2018) Challenges and Opportunities with Mobile Phone Data in Official Statistics, Conference of Statistics Stakeholders
5. G.  [D'Acquisto](https://www.amazon.it/s/ref%3Ddp_byline_sr_book_1?ie=UTF8&field-author=Giuseppe+D%27Acquisto&search-alias=stripbooks) and M. [Naldi](https://www.amazon.it/s/ref%3Ddp_byline_sr_book_2?ie=UTF8&field-author=Maurizio+Naldi&search-alias=stripbooks) Big data e privacy by design Giappichelli Editore
6. De Montjoye, Y. A., Hidalgo, C. A., Verleysen, M., & Blondel, V. D. (2013). Unique in the crowd: The privacy bounds of human mobility. Scientific reports, 3, 1376.
7. Regulation (EU) 2016/679 of the European Parliament and of the Council, dated 27 April 2016, <https://ec.europa.eu/info/law/law-topic/data-protection/reform/what-personal-data_en>
8. Pellungrini, R., Pappalardo, L., Pratesi, F., & Monreale, A. (2017, September). Fast estimation of privacy risk in human mobility data. In International Conference on Computer Safety, Reliability, and Security (pp. 415-426). Springer, Cham.
9. Samarati, P., & Sweeney, L. (1998, June). Generalizing data to provide anonymity when disclosing information. In PODS (Vol. 98, p. 188).