Longitudinal collection and analysis of mobile phone data with local differential privacy*

Héber H. Arcolezi[1][0000−0001−8059−7094], Jean-François Couchot[1][0000−0001−6437−5598], Bechara Al Bouna[2][0000−0002−7741−9905], and Xiaokui Xiao[3][0000−0003−0914−4580]

[1] Femto-ST Institute, Univ. Bourg. Franche-Comté, UBFC, CNRS, Belfort, France
{heber.hwang_arcolezi,jean-francois.couchot}@univ-fcomte.fr
[2] TICKET Lab., Antonine University, Hadat-Baabda, Lebanon
bechara.albouna@UA.EDU.LB
[3] School of Computing, National University of Singapore, Singapore
xkxiao@nus.edu.sg

Abstract. Longitudinal studies of human mobility could allow an understanding of human behavior on a vast scale. Mobile phone data call detail records (CDRs) have emerged as a prospective data source for such an important task. Nevertheless, there are significant risks when it comes to collecting this type of data, as human mobility has proven to be quite unique. Because CDRs are produced through the connection of mobile phones with mobile network operators' (MNOs) antennas, it means that users cannot sanitize their data. Once MNOs intend to use such a data source for human mobility analysis, data protection authorities such as the CNIL (in France) recommends that data be sanitized on the fly instead of collecting raw data and publishing private output at the end of the analysis. Local differential privacy (LDP) mechanisms are currently applied during the process of data collection to preserve the privacy of users. In this paper, we propose an efficient privacy-preserving LDP-based methodology to collect and analyze multi-dimensional data longitudinally through mobile connections. In our proposal, rather than regarding users as unique IDs, we propose a generic scenario where one directly collects users' sensitive data with LDP. The intuition behind this is collecting generic values, which can be generated by many users (*e.g.*, gender), allowing a longitudinal study. As we show in the results, our methodology is very appropriate for this scenario, achieving accurate frequency estimation in a multi-dimensional setting while respecting some major recommendations of data protection authorities such as the GDPR and CNIL.

Keywords: Local differential privacy · Call detail records · Mobility analytics · Multi-dimensional data · Mobile phone data.

* This work was supported by the Region of Bourgogne Franche-Comté CADRAN Project and by the EIPHI-BFC Graduate School (contract "ANR-17-EURE-0002"). The authors would also like to thank the Orange Application for Business team for their useful feedback and comments. Computations have been performed on the supercomputer facilities of "Mésocentre de Calcul de Franche-Comté".

1 Introduction

Currently, with the increasing of massive data generated by mobile phones, the acquisition of these data has attracted considerable attention. When users make a call, send SMS, or connect to the internet, a call detail record (CDR) is generated with information on users' ID, the antennas that handled the communication service (coarse level location), and the duration and type of communication, for example. CDRs are stored by mobile network operators (MNOs) for billing and legal purposes, which implies an offline, archived, and constant update of the data without changing nor deleting old records. In other words, CDRs are of easy access and, therefore, have become one of the most used data for research [6,17,10], *e.g.*, on human mobility.

In addition to CDRs, MNOs store subscription data from clients such as gender, date of birth (age), and invoice address. Such a combination of personal data makes mobile phone CDRs a rich source of information [6], which could allow research progress and improve individuals' life. For instance, CDRs can be used to model human mobility for tourism [16], to improve urban planning, and to help governmental decisions, which could be in the short-term, *e.g.*, response to natural disasters [21], the spread of new diseases/pandemics [26,36,28] like the ongoing COVID-19 outbreak [31]. Or to the long-term, to building new hospitals, schools, and improving transportation systems [23,17,27]. CDRs are also the type of data that we focus on in this paper, which will be used equivalently when mentioning mobile phone data.

1.1 Context of the problem

However, mobility data are quite sensitive as human mobility has proven to be highly unique and predictable [24]. If users can be tracked away by their presence or location, in some cases the users' home/work addresses, their religion, and habits can be disclosed. Additionally, even though MNOs have the right and duty to store CDRs, according to some privacy legislation, such as the GDPR (General Data Protection Regulation) [14], it does not mean MNOs have the right to use those data for other purposes. Therefore, once MNOs intend to use such a data source for human mobility analysis, these data must be properly sanitized. Therefore, *it is vital to improving privacy-preserving techniques to collect mobile phone data* [23,38].

For instance, one could think of a straightforward solution to publishing mobility analysis according to all CDRs collected in a given time interval (*e.g.*, a day) using k-anonymity [30] or differential privacy (DP) [11,12], which are two well-known privacy models. On the one hand, k-anonymity assumes that a trusted curator holds users' raw data and relies on hiding each released record in a crowd of $k-1$ other similar ones. However, k-anonymity might not be sufficient to guarantee users' privacy [38] since it does not offer strong guarantees and may be vulnerable to intersecting and/or homogeneity attacks, for example. On the other hand, DP assumes that a trusted curator holds the raw data from users and adds noise to output private queries. By 'trusted', we mean that curators do

not misuse or leak private information from individuals. However, this assumption does not always hold in real-life. To address non-trusted services in DP, authors in [19] introduced the concept of local differential privacy (LDP), which proposes to sanitize each individual's data independently during the process of data collection.

Indeed, DP- or k-anonymity-based approaches normally require the entire dataset of raw CDRs, which is not in accordance with recommendations of data protection authorities such as the CNIL (National Commission on Informatics and Liberty, in English) [8], in France, which requires that data be sanitized on the fly. More specifically, according to CNIL and GDPR, there are some important recommendations to be respected once MNOs aims to collect CDR-based data for analyzing human mobility. These recommendations are briefly described in the following: (i) CDRs are produced through the connection of mobile phones with carrier networks (*e.g.*, receiving or making calls). This means that users cannot sanitize their data via a built-in application, for example. Hence, all sanitization procedures must be done on the fly instead of collecting raw data and publishing private output at the end of the analysis. (ii) MNOs cannot store or work with data containing users' unique identifiers (IDs) or a hashed version of it as they are still unique IDs. For instance, if one can detect the same *hash(ID)* for many days, it violates the privacy of this user as s/he can be easily tracked away. And (iii), as a longitudinal study, users can generate multiple data points (*e.g.*, they can receive/make calls every day for a month). In the literature, this problem is referred to as *continuous observation* [13,9], and it must be addressed to avoid 'averaging' attacks.

1.2 Purpose and contributions

In summary, we describe the targeted problem of this work as follows. *How can MNOs, which store CDRs for billing purposes, also collect longitudinal and multi-dimensional data based on mobile connections to publish high-utility analysis of human mobility?* To this end, the privacy-preserving methodology must comply with the three (i)-(iii) aforementioned recommendations according to the GDPR and CNIL.

We describe our system model in the following. The first entity, namely *users*, refers to n clients $\mathcal{U} = \{u_1, u_2, ..., u_n\}$ of the MNO. The second entity is the MNOs themselves, which legally store the signature data of their clients (date of birth, gender, invoice address, ...) and billing data (CDRs) into a data server. In order to avoid data breaches and to be compliant with privacy legislation such as the GDPR, these data must be stored in a secure environment. Therefore, in the data privacy context, when users connect to the MNO's antennas, there is a need for an efficient approach for privately 'exporting' (collecting) data for high-utility mobility analytics.

Without loss of generality, we assume that all the data MNOs will extract from this data server, *for publishing human mobility analysis*, is categorical. That is, in this paper, we aim to address the aforementioned problem as a *general frequency estimation in categorical data*. The reason behind this is the collection

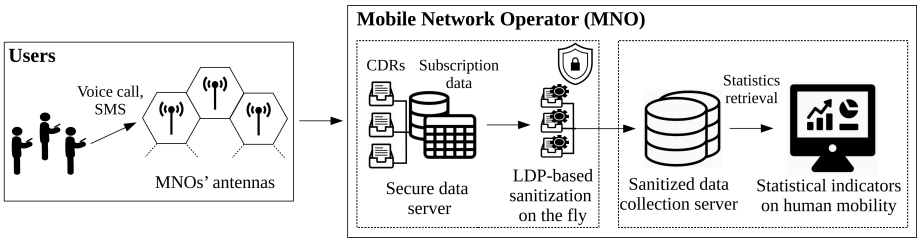

Fig. 1: Overview of our system model with an LDP-based privacy-preserving solution to sanitize users' data on the fly.

of generic values on a population through mobile connections in a given time interval to produce flows and mobility indicators based on socio-demographic data (*e.g.*, such as in [5,27]). Indeed, multi-dimensional human mobility analyses would provide more insights into people's mobility behavior by attribute values (*e.g.*, gender, origin, age-ranges). For instance, local authorities and/or organizations could take advantage of such knowledge to identify strategies to propose better decision-making solutions to society, *e.g.*, the spread of diseases, urban planning, natural disasters, and so on [10,29,20,36,26,28,21].

Therefore, in this paper, we introduce an LDP-based privacy-preserving methodology to sanitize each CDR-based data on the fly. The main reason behind our choice to use the local model is because neither MNOs nor 'trusted curators' (*e.g.*, researchers) can analyze human mobility through raw CDRs data, according to privacy legislation such as the GDPR and CNIL (in France). Fig. 1 illustrates an overview of our system model, which includes our proposed LDP-based methodology as a solution to collecting data through mobile connections.

Each time a user makes a call, or sends SMS, or connects to the internet, a CDR is generated and is stored offline for billing and legal purposes in a secure data server. Notice that the secure environment is also responsible for applying the LDP sanitization to all users' data with ϵ-LDP, where ϵ is a public parameter, before sending them to a sanitized data server. Therefore, the sanitized data server can store and aggregate this data for producing statistics through frequency estimation, which depends on the LDP algorithm and public parameter ϵ. In this scenario, both users and MNOs are safeguarded as no raw data will be collected for the purpose of human mobility analysis. On the one hand, ϵ-LDP values are robust to post-processing and can be stored and shared with 'trusted' curators for studying human mobility. However, ϵ-LDP values must not be regarded as unique identifiers. In a worst-case scenario, if one can detect a unique ϵ-LDP value for many days, it would violate the privacy of this user as s/he could be easily tracked away. In this case, we propose to use the generalized randomized response (GRR) [34,33,18] LDP mechanism, which corresponds to the situation where no particular encoding is chosen. In other words, with GRR, ϵ-LDP private reports are generic to many users (*e.g.*, feminine or masculine, for the gender attribute), which could have been generated by any user u_1 or

u_2 and, therefore, allowing a longitudinal collection of data. To summarize, this paper makes the following contributions:

- We introduce an LDP-based privacy-preserving methodology to sanitize multi-dimensional CDR-based data on the fly for two longitudinal data collection scenarios:
  - I *Frequency estimates:* In this case, MNOs initialize a single server to collect sanitized data for each time interval. This allows MNOs to publish the frequency of users per d attributes. For example, the frequency of users in a given city and day per gender, location (antenna coverage areas), age-ranges, and so on.
  - II *Cumulative frequency estimates:* In addition to point *1.*, MNOs can privately extend the analysis to include users data on different sanitized data servers. This would allow MNOs to estimate the number of users at the intersection or union of different time intervals. In this paper, we only focus on this scenario as the *Frequency estimates* scenario is already part of the latter.
- We extended the analytical analysis of GRR for multi-dimensional studies where we prove that sending only $r = 1$ attribute out of d possible ones provides *much higher utility*.
- We present extensive experiments with a 7-days, multi-dimensional, and real-world mobility dataset (CDR-driven) from [5] to validate our proposals.

Roadmap. The remainder of this paper is organized as follows. In Section 2 we review related work. In Section 3, we briefly describe the privacy notions that we are considering, *i.e.*, LDP and the GRR mechanism with its extension for multi-dimensional data. Next, we explain our proposed methodology in Section 4. In Section 5, we present our results and its discussions. Last, in Section 6 we present the concluding remarks and future directions.

2 Related Work

Numerous previous [20,21,36,16] and current studies [10,26,28,29,27] have shown the validity and benefits of mobility studies based on CDRs. Although event-based (*i.e.*, data are only available when using the MNOs' services) and with coarse location (*i.e.*, antennas that handled the service), CDRs are a real marker of human presence. Such a data source has been proven as a prominent way to analyze large-scale human mobility due to the high penetration rates of cell phones and low collection costs [6,17,10].

Regarding human mobility, some studies have shown that humans follow particular patterns with a high probability of predictability. This motivates, even more, a conscientious use of mobile phone data for allowing research progress, which could benefit individuals and society [24,23]. Indeed, in recent contexts, Oliver et al. [26] highlight the importance of mobile phone data like CDRs for

fighting the current COVID-19 pandemic [31], and [10] review and discuss applications, opportunities, and some key challenges to use CDRs for urban climate change adaption.

Concerning CDRs sanitization, the authors in [1] describe a differentially private scheme to release the spatio-temporal density of Paris regions using CDRs of about 2 million users over one week. The authors propose data pre-processing techniques such as sub-sampling and clustering, which aims at enhancing the utility of the DP mechanism. Zang and Bolot [38] have performed extensive experiments showing that the anonymization of location data from CDRs using k-anonymity leads to privacy risks. Mir et al. [22] extended a framework namely WHERE and proposed DP-WHERE, which produces synthetic CDRs to model mobility patterns of metropolitan populations respecting DP guarantees. In [3], the authors applied a DP mechanism namely BLIP (Bloom-than-fLIP) on a CDRs dataset. Each BLIP stores the users' ID, and in the end, each bit is flipped with a given probability to guarantee DP for each user. Afterward, in [2], authors proposed an upgraded version of BLIP namely Pan-Private BLIP (PP-BLIP), which guarantees privacy protection to the internal state while the BLIP is being built as well as to its output. The authors applied PP-BLIP to human mobility modeling via wi-fi connections while highlighting its extension to CDR-based data.

Notice that the aforementioned works provide a means to model human mobility through CDRs. However, when using DP, these works collect raw data and sanitize it in the end. Secondly, using BLIP or PP-BLIP would require an unbounded privacy budget as each Bloom Filter (*e.g.*, per day) uses ϵ-DP. By the sequential composition theorem [12,39], it is clear that these methods do not allow a long-term analysis. Last, these approaches collect users' ID or a hashed version of it and do not allow collecting other data rather than the presence of users itself. In this paper, we introduce a way to address these concerns with LDP, which allows a longitudinal collection of data with high privacy guarantees to each user. The local DP model has received considerable attention in both academia [32,25,19,33,34,4,15] and practical deployment [13,9] since it does not rely on sharing raw data anymore, which has a clear connection to the concept of randomized response [35]. We refer the interested reader to the survey work on LDP from Xiong et al. [37] for more insights about this approach.

3 Background

In this section, we briefly present the privacy concepts considered in this work, *i.e.*, local differential privacy (Subsection 3.1). Further, we describe the generalized randomized response LDP mechanism (Subsection 3.2), and its extension to a multi-dimensional setting based on random sampling (Subsection 3.3).

3.1 Local differential privacy (LDP)

The centralized DP outlined in the introduction assumes trusted data collectors. By 'trusted', it means that they do not steal or leak private information

from individuals. However, this assumption does not always hold in real-life. To address non-trusted services, local differential privacy [19] was proposed to preserve users' privacy in the process of data collection. Rather than trusting in a trusted curator to have the raw data and sanitize it to output queries, LDP allows users to sanitize their data before sending it to the data collector server. A randomized algorithm $\mathcal{A}$ is said to provide ϵ-local differential privacy if, for any two input values $v_1, v_2 \in Domain(\mathcal{A})$ and any possible output y of $\mathcal{A}$:

$$\Pr[\mathcal{A}(v_1) = y] \leq e^{\epsilon} \cdot \Pr[\mathcal{A}(v_2) = y]. \tag{1}$$

Intuitively, ϵ-LDP guarantees that an attacker can not distinguish whether the true value is $v1$ or $v2$ (input) with high confidence (controlled by ϵ) irrespective of the background knowledge one has. That is because both have approximately the same probability to generate the same sanitized output.

3.2 Generalized randomized response (GRR)

Randomized response (RR) was proposed by Warner [35] in 1965 to collect statistics on sensitive topics while guaranteeing the survey respondents with strong deniability. The RR model was designed for binary attributes ($v = 2$ values). Further, an LDP mechanism was introduced as a general version of the RR technique to be used for $v \geq 2$ values. This LDP mechanism is referred to as k-RR in [18], as GRR in [34], and direct encoding protocol in [33]. In this paper, we refer to this LDP mechanism as GRR.

Let $V = \{v_1, v_2, ..., v_j\}$ be a set of j values of the personal data in consideration (*e.g.*, age ranges) and ϵ be the privacy budget. Given a value v_i, *GRR(v_i)* outputs the true value v_i with probability $p = \frac{e^{\epsilon}}{e^{\epsilon}+j-1}$, and any other value v_k for $k \neq i$ with probability $q = \frac{1-p}{j-1} = \frac{1}{e^{\epsilon}+j-1}$, where $j = |V|$. A statistical-based (SB) approach for estimating the number of times $\hat{T}$ that a value v_i occurs for $i \in [1, j]$, is computed as:

$$\hat{T}(v_i) = \frac{N_i - nq}{p - q}, \tag{2}$$

where N_i is the number of times the value v_i has been reported and n is the total number of users. Wang et al. [34,33] proved that this estimator is unbiased and its variance is computed as:

$$Var[\hat{T}_{GRR}(i)] = n \cdot \frac{e^{\epsilon} + j - 2}{(e^{\epsilon} - 1)^2}, \tag{3}$$

where [34,33] remark that this variance is linear in n and j. That is, if an attribute has too many values the accuracy of using GRR decreases. Indeed, as j increases the probability p of reporting the true value decreases. Moreover, in [33], the authors have proven that if $j < 3e^{\epsilon} + 2$, the GRR mechanism presents higher utility in comparison with many other state-of-the-art LDP mechanisms, *e.g.*, Basic One-time RAPPOR [13] and optimized unary encoding [33] (OUE).

3.3 Collecting multi-dimensional data with GRR

There are few works for collecting multi-dimensional data with LDP based on random sampling [32,25,33], which mainly focused on numerical data [37]. This technique reduces both dimensionality and communication costs, which will also be the focus of this paper by extending the analysis to the GRR mechanism. Suppose there are $d \geq 2$ attributes, n users, and a privacy budget $\epsilon > 0$. A naive solution is splitting the privacy budget (*M1*), *i.e.*, assigning ϵ/d for each attribute. The other solution is based on randomly sampling (without replacement) only r attribute(s) out of d possible ones (*M2*), *i.e.*, assigning ϵ/r per attribute. Notice that both solutions satisfy ϵ-LDP according to the sequential composition theorem [12,39].

As the number of users n differs (n and rn/d) in the two solutions, we normalize the estimator $\hat{T}$ in Eq. (2) to the range 0 to 1. The GRR variance in Eq. (3) is now described as $Var[\hat{T}_{GRR}(i)/n] = \frac{1}{n} \cdot \frac{e^{\epsilon}+j-2}{(e^{\epsilon}-1)^2}$. For the first case (*M1*), the GRR variance is $Var_1 = \frac{1}{n} \cdot \frac{e^{\epsilon/d}+j-2}{(e^{\epsilon/d}-1)^2}$ and for the second case (*M2*), the GRR variance is $Var_2 = \frac{d}{nr} \cdot \frac{e^{\epsilon/r}+j-2}{(e^{\epsilon/r}-1)^2}$. The objective is finding r, which minimizes Var_2 and guarantees that $Var_2 < Var_1$.

Following the work in [32], we multiply Var_2 by ϵ. Next, let $x = r/\epsilon$ be the independent variable and Var_2 rewritten as $y = \frac{1}{x} \cdot \frac{e^{1/x}}{(e^{1/x}-1)^2}$ be the dependent one; d/n and $j-2$ are omitted as they are simply summing or multiplication factors. Since y is an increasing function, *i.e.*, y increases as the x value increases, we then have a minimum and optimal when $r = 1$. The remaining question is if $Var_1 - Var_2 > 0$. Since

$$Var_1 - Var_2 = \frac{1}{n}\left(\frac{e^{\epsilon/d}}{(e^{\epsilon/d}-1)^2} - \frac{d \cdot e^{\epsilon}}{(e^{\epsilon}-1)^2}\right), \tag{4}$$

omitting $j-2$, it has been proven in [33] that $Var_1 - Var_2$ is always positive, and hence, the proof ends.

In this multi-dimensional setting (*M2*), one can see our solution as applying the GRR mechanism only once in d times ($1/d$) and reporting nothing $1 - 1/d$ times. In short, Alg. 1 shows the pseudocode of using GRR for multi-dimensional data collection, which will be referred to as GRR-*M2* for the rest of this paper. Given the set $A = \{A_1, ..., A_d\}$ of all d attributes and a tuple $t = [v_1, ..., v_d]$ with the user's true values, the GRR-*M2* algorithm returns a tuple $t^* = \langle r, GRR(v_r)\rangle$, *i.e.*, with the sampled attribute r and its ϵ-LDP value. Notice that, to ensure (strengthen) users' privacy over time, each user must always report the same unique attribute r. On the server-side, the SB estimator in Eq.(2) to the number of times $\hat{T}$ that a value v_i occurs for $i \in [1, j]$ has to be scaled d times.

Algorithm 1 GRR for multi-dimensional data collection (GRR-*M2*)

Input : tuple $t = [v_1, ..., v_d]$, set $A = \{A_1, ..., A_d\}$, and ϵ.
Output : tuple $t^* = \langle r, GRR(v_r) \rangle$.
$r \leftarrow Uniform(\{1, 2, ..., d\})$
$B \leftarrow t[r] = v_r$
$b \leftarrow Bern\left(e^{\epsilon}/(e^{\epsilon} + |A_r| - 1)\right)$
if $b = 1$**:**
    $B' = v_r$
else:
    $B' \leftarrow Uniform(\{A_r/v_r\})$
return : $t^* = \langle r, B' \rangle$

4 LDP-Based Privacy-Preserving Longitudinal Data Collection Through Mobile Connections

In this section, according to the system overview in Fig. 1, we detail our LDP-based solution (Subsection 4.1) regarding the *Cumulative frequency estimates* scenario outlined in the introduction and its limitations (Subsection 4.2). Notice that the *Frequency estimates* scenario of collecting data only per single time intervals is part of the *Cumulative frequency estimates* scenario, and one can intuitively simplify some steps to apply only it.

4.1 Proposed methodology

Fig. 2 illustrates the overview of our methodology applied to collect users' data for days and union of days in a flow chart. Without loss of generality, we present our methodology for days, but it can be extended to any timestamp one desires as users' LDP data are generic to many users by using the GRR mechanism.

1. **Initialization.** According to the left side of Fig. 2, the secure-server defines the privacy guarantee ϵ, which is uniform for all users. Let Nb be the whole period of analysis (*e.g.*, total number of days), the MNOs' sanitized data server initializes $Nb(Nb+1)/2$ empty databases. For instance, if $Nb = 3$ one will have $set_{db} = \{D_1, D_2, D_2 \cup D_1, D_3, D_3 \cup D_2, D_3 \cup D_2 \cup D_1\}$.
2. **LDP-based sanitization on the fly.** The MNOs' secure-server is responsible to applying the LDP-based sanitization model on the fly (*cf.* Fig. 1). This process uses our GRR-*M2* solution (Alg. 1). Because LDP is applied on the secure server of MNOs, there are a few issues to be taken into account. To ensure and strengthen privacy over time using GRR-*M2*, users must always report the same unique attribute r. To solve this issue, we suggest that MNOs generate a random seed that will be associated with each user for a cryptographically secure pseudorandom number generator. Thanks to this, the same memoized and sanitized B' can always be assigned for each user, according to the same and unique sampled attribute r. Indeed, if different values B' were to be sent in each day, in the long-term, attackers who

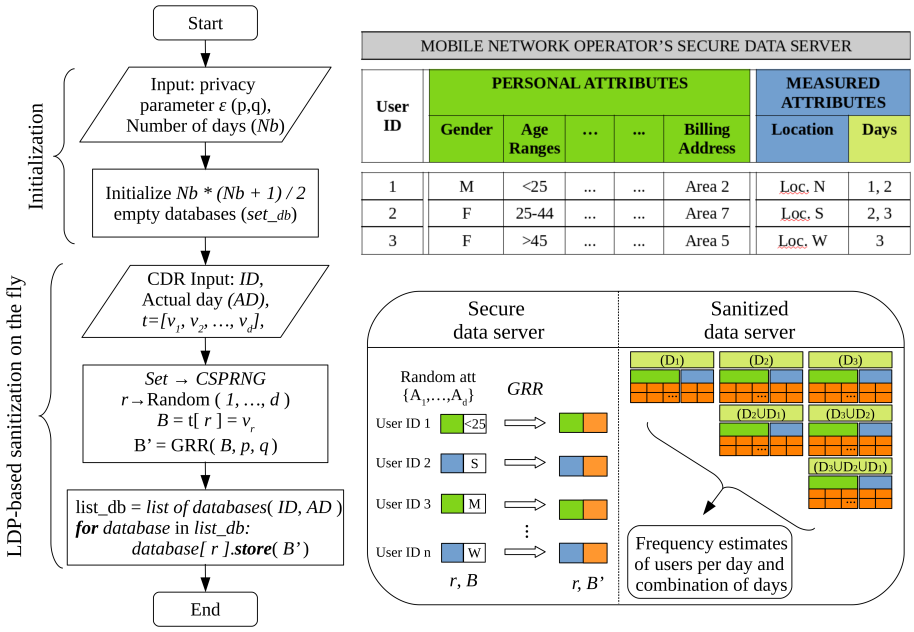

Fig. 2: Overview of our LDP-based privacy-preserving methodology to collecting data through mobile connections for days and union of days.

can isolate reports from a single user could infer with high confidence the true sensitive value by averaging attacks [13,9]. Also, in the real world, it is noteworthy noticing that in a given time interval between ts_{ini} to ts_{end}, the same user can generate multiple connections to MNOs' antennas. This de-duplication issue can be simply solved by the MNO's secure-server via Bloom filters [7], for example.

3. **Private data collection.** Each time a user connects to MNO's antennas, a CDR is generated, which contains the user's identifier (ID) and timestamp (actual day - AD), for example. CDRs are stored offline in a data server, which contains subscription data of users such as gender, date of birth, origin (invoice address), for example. Hence, without loss of generality, each user u_i $(1 \leq i \leq n)$ has a discrete-encoded tuple record $t = [v_1, v_2, ..., v_d]$, which contains d categorical attributes $A = \{A_1, A_2, ..., A_d\}$. Therefore, the MNO's secure-server consistently uses the memoized sanitized data B' of users to send for storage according to the users' list of databases $list_{db}$. That is, by knowing the days this user "was present" (by CDRs), it allows calculating $list_{db}$, which is a list of databases (days and union of days) to store the private report of the user. We later explain in an example how to calculate $list_{db}$.

4. **Generating statistics.** At the end of the analysis period, the MNO's sanitized data server should have $Nb(Nb+1)/2$ databases with only ϵ-LDP reports. On the one hand, MNOs can publish an accurate mobility scenario according to the number of reports (B') in each sanitized database. The latter represents the *exact* number of users present per day and union of days. Last, for each sanitized database, one can estimate the frequency of this population for all d attributes using the SB estimator in Eq. (2), which has to be scaled d times.

Example to calculate list of databases. To calculate the $list_{db}$ for each user, consider the right side of Fig. 2, which has data for $Nb = 3$ days. First, let Actual Day $AD = 1$ (the first day of analysis). So, user $ID = 1$ is detected by the secure-server and his $list_{db} = \{D_1, D_2 \cup D_1, D_3 \cup D_2 \cup D_1\}$. The reason behind this is that if this user does not appear anymore, we have considered his data in the whole analysis. Next, let $AD = 2$. For the same user $ID = 1$, the secure-server knows he was present in both two days, hence, his $list_{db} = \{D_2, D_3 \cup D_2\}$ as the previous day his data was already stored in $D_2 \cup D_1$ and $D_3 \cup D_2 \cup D_1$. And, for the user $ID = 2$, her $list_{db} = \{D_2, D_2 \cup D_1, D_3 \cup D_2, D_3 \cup D_2 \cup D_1\}$ to guarantee her data is considered in each past union and future ones in the case she does not show up anymore. Without loss of generality the same procedure is applied in $AD = 3$.

4.2 Limitations

The first key limitation we see in our methodology is the storage factor, which is specifically related to the *Cumulative frequency estimates* scenario, *i.e.*, collecting users' data per day and union of days. For instance, MNOs need to initialize $Nb(Nb+1)/2$ empty databases where if one wishes to analyze an enhanced detailed scenario, it grows up very fast (*i.e.*, with at least an $Nb^2/2$ factor). However, this scenario is only intended in special mobility analytics cases, *e.g.*, tourism events, natural disasters, following up spread of diseases, etc. In addition, there is high power for computation and powerful tools to deal with big data nowadays. One way to smooth this problem in, *e.g.*, daily scenarios, is to exclude the stored data after retrieving statistics.

Further, there is an important loss of information by not calculating the intersection of users through days. That is, we propose to compute the number of users per union of days as it may have very few users per intersection. The latter would not produce accurate frequency estimations due to the LDP formulation, which is data-hungry. At first glance, one can surely compute the pair-wise intersection for any two days in the analysis period using $|A \cap B| = |A| + |B| - |A \cup B|$. One possibility of solving the whole problem is to use the methodology from [5], which models our proposed mobility scenario (days and union of days) as a linear program to find a solution for all intersections. Besides, for the case where one can have sufficient data samples per pair-, triple-, ..., and Nb-wise intersections, one can easily extend our methodology for such a case. However, the storage

factor is even bigger as MNOs would have to initialize $2^{Nb} - 1$ empty databases (all combinations of days).

Last, the *memoization* step of always reporting the same sanitized value for the unique sampled attribute can be effective to the cases where the true client's data does not vary (static) [13,9]. On the other hand, a measured attribute such as location is dynamic. As highlighted in the literature [24], humans mobility is very predictable, which means that they follow well-defined patterns, *e.g.*, alternating between $l1$ (home), $l2$ (work), and $l3$ (*e.g.*, hobby). Yet, in our privacy-preserving architecture (Fig. 1), the collected/stored sanitized data are 'uncorrelated' from users, as no ID will be stored. Therefore, under the worst-case scenario where attackers can isolate reports from a single user whose random dimension is location, they could learn the $l = 3$ memoized ϵ-LDP values. But, they can hardly (controlled by ϵ-LDP) identify these real locations thanks to memoization.

5 Results and Discussion

In this section, we report the results obtained by applying our proposed methodology in a real-world open dataset (Subsection 5.1) and its discussions (Subsection 5.2). The codes we developed and used for all experiments are available in a Github repository[1].

Dataset. This is a longitudinal and multi-dimensional dataset of VHs [5] resulted by inferring several statistics of human mobility; as stated by the authors, statistics were generated through sanitized mobile connections (CDR-driven). We excluded the data from 'Foreign tourist' users regarding the 'Visitor category' attribute. Hence, our filtered dataset aggregates a population of $87,098$ French users per seven days with approximately $26,700$ unique users on average per day. There are $d = 6$ attributes we are interested in, which are described in the following. *Gender* is masculine or feminine. *Age* has 7 age ranges. *GeoLife* has 12 socio-professional categories. *Region* has 22 regions in France that users are billed. *Sleeping Area* represents 11 areas that users spent the night (location). And, *Visit Duration* has 10 time-ranges in which visitors were detected each day. The two last attributes are measured ones where *Sleeping Area* is static and *Visit Duration* is dynamic.

Evaluation and metrics. We vary the *total* privacy budget ϵ in the range $[0.5, 1, 2, 3, 4, 5, 6]$ to evaluate the privacy-utility trade-off. We use the root mean square error (RMSE) metric to measure our results.

Setup: Let $Nb = 7$ days to be the whole analysis period, we then have $Nb(Nb+1)/2 = 28$ databases considering each day and union of days combination as $set_{db} = \{D_1, ..., D_3 \cup D_2 \cup D_1, ..., D_7 \cup D_6 \cup ... \cup D_1\}$. Notice that, at the same time, we can empirically evaluate the privacy-utility trade-off according to data size, *i.e.*, each day has around $26,700$ unique users, while the last union of days $D_7 \cup D_6 \cup ... \cup D_1$ has all $87,098$ users.

[1] https://github.com/hharcolezi/ldp-protocols-mobility-cdrs

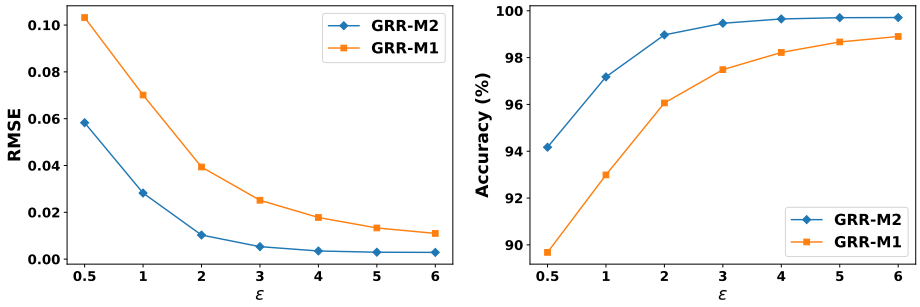

Fig. 3: Average RMSE and accuracy (y-axis) VS privacy budget ϵ (x-axis) analysis for our GRR-*M2* solution and the GRR-*M1* naive one.

Comparing methodologies. We consider for evaluation the following approaches:

- Our GRR-*M2* solution of using GRR in a multi-dimensional setting, which samples a unique attribute r and assigns the whole privacy budget ϵ to it;
- A naive solution of using GRR in a multi-dimensional setting, which splits the privacy budget among d attribute, *i.e.*, ϵ/d (GRR-*M1*).

5.1 Cumulative frequency estimates results

Fig. 3 illustrates the average RMSE values (y-axis) varying the privacy budget ϵ (x-axis) regarding all databases in set_{db} for our GRR-*M2* solution and the GRR-*M1* naive one (left-side); and, the corresponding $accuracy = 1 - RMSE$, which is rather intuitive of how much privacy budget to use for achieving a given "accuracy" (right-side). In more detail, Fig. 4 illustrates for both methods the RMSE results (y-axis) according to the privacy budget ϵ for each day and union of days (x-axis), *e.g.*, '321' refers to $D_3 \cup D_2 \cup D_1$. Without loss of generality, we excluded $\epsilon = 0.5$ in Fig. 4 to improve the visibility of the other curves. Finally, for the sake of illustration, Fig. 5 exhibits the frequency estimation results for a single day (D_7) and for the union of all days ($D_7 \cup D_6 \cup ... \cup D_1$) using our GRR-*M2* and $\epsilon = 1$.

5.2 Discussion

As one can notice in the results, our LDP-based methodology can be well applied to the longitudinal collection and analysis of multi-dimensional data for human mobility modeling. It is worthy noticing that we selected the GRR mechanism because the set of possible values in a given attribute is kept at j. If one-hot-encoding is used as in the Basic One-time RAPPOR mechanism [13] or OUE [33], each bit can be either 0 or 1, which means that the set of possible reports for an attribute with j values has now 2^j possibilities. This is highly important to take

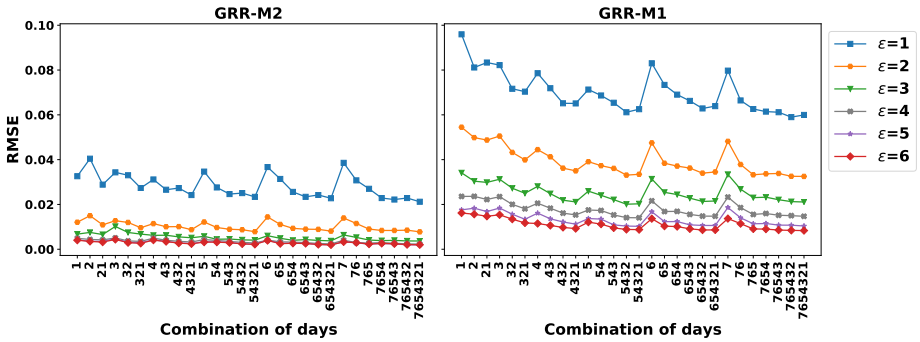

Fig. 4: RMSE (y-axis) analysis varying the privacy budget ϵ considering each combination of days (x-axis). Notice that these are discrete values. However, curves are drawn to ease interpretation.

into account, as, for a longitudinal collection of data as we proposed, one has to try to ensure that ϵ-LDP values are not indirect unique identifiers. Moreover, high-utility mobility indicators can be achieved by generalizing attributes to a well-defined set of values. That is, aiming to respect $j < 3e^{\epsilon} + 2$ or at least not being too far, as GRRs' utility, depends on [33].

Overall, our GRR-*M2* consistently and considerably outperforms the baseline GRR-*M1*. In Fig. 4, except for $\epsilon = 1$, our GRR-*M2* curves are under even to the best one of GRR-*M1* using the highest privacy budget $\epsilon = 6$. As also highlighted in the literature [32,25,33], privacy budget splitting is sub-optimal, which leads to higher estimation error. Indeed, in a multi-dimensional setting, the combination of privacy budget splitting and high numbers of values j in a given attribute (*e.g.*, *Region* with 22 values) leads to lower data utility even for high privacy regimes. On the other hand, our GRR-*M2* solution based on random sampling uses the whole privacy budget to a single attribute, and this problem is, hence, minimized. However, there is also an error provided by the sampling technique, which is reduced by correctly choosing the number of attributes $r < d$ as we did for GRR in Subsection 3.3.

More precisely, our GRR-*M2* solution presents an 'accuracy' over 94% for any privacy budget tested. In Fig. 3, with $\epsilon = 1$ that is considered a good privacy-utility trade-off, while our GRR-*M2* already approaches 98% of 'accuracy', the GRR-*M1* can only get this close when $\epsilon \geq 3$ approximately. Additionally, in Figs. 4 and 5, it is noteworthy that the RMSE decreases as the data size increases. This is due to the LDP setting, which requires a big amount of data to guarantee a good balance of noise (data-hungry). In our case, single days (*e.g.*, D_7) have less data points comparing to the union of all days (*e.g.*, $D_7 \cup D_6 \cup ... \cup D_1$), and hence they are generally the peak-values in Fig. 4. However, these peak values are smoothed using our GRR-*M2*, which induce less error by sampling a single attribute for each user.

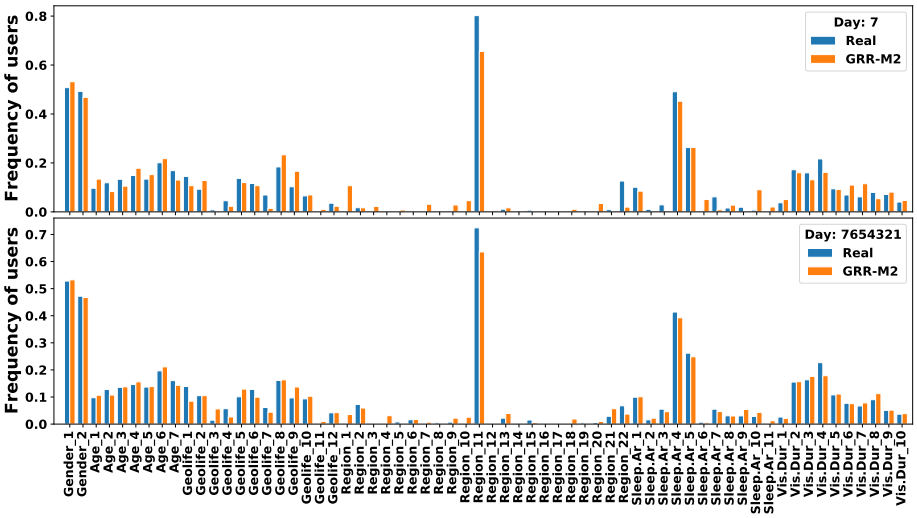

Fig. 5: Comparison between real and estimated frequencies for a single day and to the union of all days using our GRR-*M2* solution and $\epsilon = 1$.

Finally, our proposed LDP-based methodology satisfies the three (i)-(iii) recommendations of data protection authorities, described in Subsection 1.1. For instance, the MNOs' secure-server applies an LDP mechanism to sanitize all data on the fly (i) while storing no users' ID (ii). Furthermore, ϵ-LDP private reports are not (indirect) unique IDs, *i.e.*, they are generic values, which could have been generated by any user u_1 or u_2 and, thus, allowing a longitudinal collection of data (iii). Besides, each time users connect, they will always report the same $r = 1$ attribute out of d possible ones. That is, even though users appear all days in the analysis (in this dataset $\sim 0.2\%$ of users), they will never report the remaining $d - 1$ attributes, which were not directly sampled. And last, our solution also safeguards MNOs as the sanitized data server (*cf.* Fig. 1) will not collect raw or pseudonymized CDRs, for the purpose of human mobility analysis, but, rather, ϵ-LDP values that are robust to post-processing.

6 Conclusion

This work investigates the problem of longitudinally collecting and analyzing human mobility data through mobile connections. Following some major recommendations from data authorities such as the GDPR and CNIL (in France), we proposed an LDP-based privacy-preserving methodology for collecting data, on the fly, through mobile connections while providing high privacy guarantees for users. More precisely, such a privacy-preserving methodology would allow MNOs to use and/or share the sanitized data, as LDP is robust to post-processing, for publishing mobility indicators while protecting the privacy of their clients. To

this end, we extended the analysis of a state-of-the-art LDP protocol named GRR [34,33,18] for multi-dimensional studies, referred to as GRR-*M2* in this paper.

As shown in the results, the proposed LDP-based methodology using our GRR-*M2* solution is capable of collecting and calculating accurate multi-dimensional frequency estimates (*cf.* Fig. 5, for example) for human mobility modeling. Indeed, in the mobility scenario we propose, the number of users per day and union of days is exactly the number of ϵ-LDP collected reports. For future work, we suggest and intend the following directions: to experimentally validate our proposed methodology using actual data from an MNO; to investigate inference attacks to check whether individuals who report dynamic attributes, such as location, are more endangered than others who report, *e.g.*, gender; and, to design LDP mechanisms for this longitudinal and multi-dimensional task considering both numerical and categorical data.

References

1. Acs, G., Castelluccia, C.: A case study: Privacy preserving release of spatio-temporal density in paris. In: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD. ACM Press (2014). https://doi.org/10.1145/2623330.2623361
2. Alaggan, M., Cunche, M., Gambs, S.: Privacy-preserving wi-fi analytics. Proceedings on Privacy Enhancing Technologies **2018**(2), 4–26 (Apr 2018). https://doi.org/10.1515/popets-2018-0010
3. Alaggan, M., Gambs, S., Matwin, S., Tuhin, M.: Sanitization of call detail records via differentially-private bloom filters. In: Data and Applications Security and Privacy XXIX, pp. 223–230. Springer International Publishing (2015)
4. Alvim, M., Chatzikokolakis, K., Palamidessi, C., Pazii, A.: Invited paper: Local differential privacy on metric spaces: Optimizing the trade-off with utility. In: 2018 IEEE 31st Computer Security Foundations Symposium (CSF). IEEE (Jul 2018). https://doi.org/10.1109/csf.2018.00026
5. Arcolezi, H.H., Couchot, J.F., Baala, O., Contet, J.M., Bouna, B.A., Xiao, X.: Mobility modeling through mobile data: generating an optimized and open dataset respecting privacy. In: 2020 International Wireless Communications and Mobile Computing (IWCMC). IEEE (Jun 2020). https://doi.org/10.1109/iwcmc48107.2020.9148138
6. Blondel, V.D., Decuyper, A., Krings, G.: A survey of results on mobile phone datasets analysis. EPJ Data Science **4**(1) (Aug 2015). https://doi.org/10.1140/epjds/s13688-015-0046-0
7. Broder, A., Mitzenmacher, M.: Network applications of bloom filters: A survey. Internet Mathematics **1**(4), 485–509 (Jan 2004). https://doi.org/10.1080/15427951.2004.10129096
8. CNIL: Commission nationale de l'informatique et des libertés. https://www.cnil.fr/en/home (1978), online; accessed 10 May 2020
9. Ding, B., Kulkarni, J., Yekhanin, S.: Collecting telemetry data privately. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30, pp. 3571–3580. Curran Associates, Inc. (2017)

10. Dujardin, S., Jacques, D., Steele, J., Linard, C.: Mobile phone data for urban climate change adaptation: Reviewing applications, opportunities and key challenges. Sustainability **12**(4), 1501 (Feb 2020). https://doi.org/10.3390/su12041501
11. Dwork, C.: Differential privacy. In: Automata, Languages and Programming, pp. 1–12. Springer Berlin Heidelberg (2006). https://doi.org/10.1007/11787006_1
12. Dwork, C., Roth, A., et al.: The algorithmic foundations of differential privacy. Foundations and Trends® in Theoretical Computer Science **9**(3–4), 211–407 (2014)
13. Erlingsson, U., Pihur, V., Korolova, A.: Rappor: Randomized aggregatable privacy-preserving ordinal response. In: Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security. ACM, New York, NY, USA (2014)
14. European-Commission: 2018 reform of EU data protection rules. https://gdpr-info.eu/ (2018), online; accessed 10 April 2020
15. Fernandes, N., Lefki, K., Palamidessi, C.: Utility-preserving privacy mechanisms for counting queries. In: Models, Languages, and Tools for Concurrent and Distributed Programming, pp. 487–495. Springer International Publishing (2019). https://doi.org/10.1007/978-3-030-21485-2_27
16. Heerschap, N., Ortega, S., Priem, A., Offermans, M.: Innovation of tourism statistics through the use of new big data sources. In: 12th global forum on tourism statistics, Prague, CZ. vol. 716 (2014)
17. Jacques, D.C.: Mobile phone metadata for development. arXiv preprint arXiv:1806.03086 (2018)
18. Kairouz, P., Bonawitz, K., Ramage, D.: Discrete distribution estimation under local privacy. arXiv preprint arXiv:1602.07387 (2016)
19. Kasiviswanathan, S.P., Lee, H.K., Nissim, K., Raskhodnikova, S., Smith, A.: What can we learn privately? In: 2008 49th Annual IEEE Symposium on Foundations of Computer Science. IEEE (Oct 2008). https://doi.org/10.1109/focs.2008.27
20. Kishore, N., Mitchell, R., Lash, T.L., Reed, C., Danon, L., Sigmundsdóttir, G., Vigfusson, Y.: Flying, phones and flu: Anonymized call records suggest that keflavik international airport introduced pandemic H1N1 into iceland in 2009. Influenza and Other Respiratory Viruses **14**(1), 37–45 (Nov 2019). https://doi.org/10.1111/irv.12690
21. Lu, X., Bengtsson, L., Holme, P.: Predictability of population displacement after the 2010 haiti earthquake. Proceedings of the National Academy of Sciences **109**(29), 11576–11581 (Jun 2012). https://doi.org/10.1073/pnas.1203882109
22. Mir, D.J., Isaacman, S., Caceres, R., Martonosi, M., Wright, R.N.: DP-WHERE: Differentially private modeling of human mobility. In: 2013 IEEE International Conference on Big Data. IEEE (Oct 2013). https://doi.org/10.1109/bigdata.2013.6691626
23. de Montjoye, Y.A., Gambs, S., Blondel, V., Canright, G., de Cordes, N., Deletaille, S., Engø-Monsen, K., Garcia-Herranz, M., Kendall, J., Kerry, C., Krings, G., Letouzé, E., Luengo-Oroz, M., Oliver, N., Rocher, L., Rutherford, A., Smoreda, Z., Steele, J., Wetter, E., Pentland, A.“., Bengtsson, L.: On the privacy-conscientious use of mobile phone data. Scientific Data **5**(1) (Dec 2018). https://doi.org/10.1038/sdata.2018.286
24. de Montjoye, Y.A., Hidalgo, C.A., Verleysen, M., Blondel, V.D.: Unique in the crowd: The privacy bounds of human mobility. Scientific Reports **3**(1) (Mar 2013). https://doi.org/10.1038/srep01376
25. Nguyên, T.T., Xiao, X., Yang, Y., Hui, S.C., Shin, H., Shin, J.: Collecting and analyzing data from smart device users with local differential privacy. ArXiv **abs/1606.05053** (2016)

26. Oliver, N., Lepri, B., Sterly, H., Lambiotte, R., Deletaille, S., Nadai, M.D., Letouzé, E., Salah, A.A., Benjamins, R., Cattuto, C., Colizza, V., de Cordes, N., Fraiberger, S.P., Koebe, T., Lehmann, S., Murillo, J., Pentland, A., Pham, P.N., Pivetta, F., Saramäki, J., Scarpino, S.V., Tizzoni, M., Verhulst, S., Vinck, P.: Mobile phone data for informing public health actions across the COVID-19 pandemic life cycle. Science Advances **6**(23), eabc0764 (Apr 2020). https://doi.org/10.1126/sciadv.abc0764
27. Orange-Business-Services: Flux vision: real time statistics on mobility patterns. https://www.orange-business.com/en/products/flux-vision (2013), online; accessed 1 Jul 2020
28. Pollina, E., Busvine, D.: European mobile operators share data for coronavirus fight. https://www.reuters.com/article/us-health-coronavirus-europe-telecoms-idUSKBN2152C2 (2013), online; accessed 1 Dec 2020
29. Rhoads, D., Serrano, I., Borge-Holthoefer, J., Solé-Ribalta, A.: Measuring and mitigating behavioural segregation using call detail records. EPJ Data Science **9**(1) (Mar 2020). https://doi.org/10.1140/epjds/s13688-020-00222-1
30. Sweeney, L.: k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems **10**(05), 557–570 (Oct 2002). https://doi.org/10.1142/s0218488502001648
31. Wang, C., Horby, P.W., Hayden, F.G., Gao, G.F.: A novel coronavirus outbreak of global health concern. The Lancet **395**(10223), 470–473 (Feb 2020). https://doi.org/10.1016/s0140-6736(20)30185-9
32. Wang, N., Xiao, X., Yang, Y., Zhao, J., Hui, S.C., Shin, H., Shin, J., Yu, G.: Collecting and analyzing multidimensional data with local differential privacy. In: 2019 IEEE 35th International Conference on Data Engineering (ICDE). IEEE (Apr 2019)
33. Wang, T., Blocki, J., Li, N., Jha, S.: Locally differentially private protocols for frequency estimation. In: 26th USENIX Security Symposium (USENIX Security 17). pp. 729–745. USENIX Association, Vancouver, BC (Aug 2017)
34. Wang, T., Li, N., Jha, S.: Locally differentially private frequent itemset mining. In: 2018 IEEE Symposium on Security and Privacy (SP). IEEE (May 2018). https://doi.org/10.1109/sp.2018.00035
35. Warner, S.L.: Randomized response: A survey technique for eliminating evasive answer bias. Journal of the American Statistical Association **60**(309), 63–69 (Mar 1965). https://doi.org/10.1080/01621459.1965.10480775
36. Wesolowski, A., Buckee, C.O., Bengtsson, L., Wetter, E., Lu, X., Tatem, A.J.: Commentary: Containing the ebola outbreak - the potential and challenge of mobile network data. PLoS Currents (2014). https://doi.org/10.1371/currents.outbreaks.0177e7fcf52217b8b634376e2f3efc5e
37. Xiong, X., Liu, S., Li, D., Cai, Z., Niu, X.: A comprehensive survey on local differential privacy. Security and Communication Networks **2020**, 1–29 (Oct 2020). https://doi.org/10.1155/2020/8829523
38. Zang, H., Bolot, J.: Anonymization of location data does not work. In: Proceedings of the 17th annual international conference on Mobile computing and networking - MobiCom. ACM Press (2011). https://doi.org/10.1145/2030613.2030630
39. Zhu, T., Li, G., Zhou, W., Yu, P.S.: Differential Privacy and Applications. Springer International Publishing (2017). https://doi.org/10.1007/978-3-319-62004-6